Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Anotação semi-automática baseada em ontologia, busca e relacionamento semântico entre textos: Proposta para um sistema de gerenciamento de conteúdo

Vitor Silva de Deus

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr. Edison Ishikawa

Brasília
2018

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Anotação semi-automática baseada em ontologia, busca e relacionamento semântico entre textos: Proposta para um sistema de gerenciamento de conteúdo

Vitor Silva de Deus

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr. Edison Ishikawa (Orientador)
CIC/UnB

Prof. Dr. Edgard Costa Oliveira
EPR/UnB

Dr. Benedito Medeiros Neto
FAC/UnB

Dr. Marcio Vitorino
FCI/UnB

Prof. Dr. Edison Ishikawa
Coordenador do Bacharelado em Ciência da Computação

Brasília, 21 de junho de 2018

Dedicatória

Dedico aos honráveis brasileiros e brasileiras que lutam e lutaram pelo ensino superior gratuito e de qualidade no Brasil.

Agradecimentos

Agradeço a toda comunidade de software livre dos quais pude reusar e tornar alcançaveis os objetivos deste trabalho. Agradeço aos autores de toda a produção científica que tive acesso gratuito na internet. Agradeço ao Professor Edison Ishikawa pela indispensável orientação.

Resumo

A semântica ainda é um desafio para a automação e para a melhoria do relacionamento entre homem e máquina. Os sistemas de gerenciamento de conteúdo apresentam várias possibilidades de melhoria com o processamento semântico. Uma redação jornalística com um sistema de gerenciamento de conteúdo que possibilite busca semântica, relacionamento semântico entre artigos jornalísticos e comunicação com sistemas semânticos externos pode incrementar a capacidade produtiva dos profissionais e facilitar a vida de quem consome o conteúdo, por exemplo. Este trabalho apresenta um protótipo funcional de um sistema de gerenciamento de conteúdo semântico com foco na construção de anotações semânticas semi-automáticas baseadas em uma ontologia de domínio, reutilizar as anotações para busca e construção de relacionamentos entre textos armazenados no sistema e prover uma interface semântica para acesso do conteúdo do sistema por sistemas externos ao sistema de informação no qual possa ser implantado. São apresentados o algoritmo de anotação, os casos de uso de anotação, criação e edição de artigos, uma abordagem para busca semântica dos mesmos e duas abordagens para construção de relacionamentos entre os textos anotados. O sistema aumenta muito a velocidade em que se pode anotar semanticamente um artigo além permitir que o usuário adicione e remova anotações que foram sugeridas automaticamente. As duas abordagens para relacionamento de textos se mostraram bastante precisas e úteis. A ferramenta de busca é muito versátil por permitir a busca em campos de dados semânticos e não semânticos ao mesmo tempo e também o uso de operadores lógicos em todos eles.

Palavras-chave: Web Semântica,Busca Semântica,Computação Semântica, Web 3.0, Anotação semântica,Sistema de Gerenciamento de Conteúdo

Abstract

Semantics is still a challenge to automation and to the improvement of the relationship between humans and machines. The content management systems have a lot of improvement possibilities using semantics. A news writing with a content management system that provides semantic search, semantic relationships between articles and communication with external semantic systems can improve the productivity of the writers and make the reader task easier. This work presents a functional prototype of a content management system that focus in the construction of semantic annotations based on domain ontology, reuse annotations in search and relationship construction between stored texts and provide a semantic interface for external systems. Are presented the annotation algorithm, the use cases of annotation, article creation and editing, an approach for for a semantic search and two approaches for build semantic relationships between texts. The system enable users two create semantic annotations quickly and allow them to remove and add annotations that were suggested or not by the annotations algorithm. The two approaches for semantic relationships between texts are accurate and useful. The search tool is versatile because it allows users to search in semantic and non semantic fields at the same time and use logic operators in all fields.

Keywords: Semantic Web,Semantic Search,Ontology,Semantic Computing, Web 3.0, Semantic Annotation,Content Management Sistem

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

CAPES Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

DSR Design Science Research.

HTML Hypertext Markup Language.

OWL Web Ontology Language.

RDF Resource Description Framework.

SPARQL Protocol and RDF Query Language.

SQL Structured Query Language.

URI Universal Resource Identifier.

Wordnet Base de dados léxica para o inglês.

XML Extensible Markup Language.

Capítulo 1

Introdução

A internet de hoje produz uma enorme massa de dados que podem ser buscados e processados para obtenção de métricas e estatísticas. Entretanto, em que nível chega esse processamento? Até que nível é possível processar o conteúdo da Web e extrair informações sobre ele? Fato é que existe muita informação que não é extraída pelo simples fato de que o conteúdo da Web não é livre de contexto e as máquinas não "entendem"contexto ou o significado do conteúdo em cada contexto. Ou seja, há frases com ambiguidades semânticas e regionalismos que contém informações que não podem ser extraídas por um computador apenas a partir do texto da frase.

Além do contexto, os conceitos possuem relações com outros conceitos que não podem ser extraídos por computadores simplesmente a partir da citação textual desses conceitos. Por exemplo, o carnaval do Rio de Janeiro está associado ao conceito Sapucaí mas não podemos extrair essa informação de dois textos que falem separadamente de cada coisa. Desse modo vemos que o processamento do conteúdo da Web atual (ou Web 2.0) a partir de documentos vai até o nível de comparação de caracteres, mineração de texto e extração de estatísticas.

A conscientização dessas limitações tem levado ao desenvolvimento de tecnologias, ferramentas e estruturas que possam possibilitar que máquinas processem o conteúdo semântico presente na Web. O conteúdo na Web que tenha estruturas pensadas para possibilitar esse processamento semântico permite a inferência e a extração de várias informações que seriam naturalmente entendidas somente por humanos. O conceito Sapucaí pode automaticamente ser relacionado a cidade do Rio de Janeiro, ao carnaval ou uma data de feriado, basta que os conceitos presentes no texto estejam atrelados a um referencial compartilhado na Web que possa ser processado por máquinas. Essa nova abordagem na construção do conteúdo na Web é conhecida como Web Semântica ou Web 3.0 [4].

O conteúdo jornalístico é uma parcela importante do conteúdo produzido na Web. Os meios físicos de publicação de notícias como jornais impressos e revistas têm dado

lugar ao conteúdo online na internet. A pouca automação existente no processamento do conteúdo semântico deixa lacunas tanto para quem lê quanto para quem produz conteúdo jornalístico.

Um jornalista ao escrever sobre *febre amarela*, por exemplo, precisa de ter referências sobre o assunto. Entretanto, se o conteúdo da Web não relaciona seus conceitos e nem existem estruturas semânticas que permitam isso, o jornalista gastará bastante tempo pesquisando. Seria ideal se ao pesquisar com a chave de busca *febre amarela*, ele obtivesse resultados de cidades que tiveram epidemias, vacinas, outros artigos, mosquito transmissor e outros. É claro que o conteúdo textual de um documento também pode ter essas informações, ou seja, um texto que tenha o segmento *aedes aegypti* pode também ter o trecho *febre amarela*, mas não necessariamente. Entretanto, na Web semântica, o mosquito está relacionado a febre amarela com um referencial compartilhado entre todos os conteúdos que o referenciam na Web.

Do lado do leitor também existem lacunas. Por exemplo, se um leitor quiser ler artigos que falem do sintoma febre, em uma busca puramente textual ele pode receber resultados de *febre amarela*, *febre chikungunya*, *febre do zika* e etc. Entretanto essas febres se referem as doenças e não ao sintoma febre.

Visto isso, percebe-se que processar conteúdo semântico com máquinas é aumentar a capacidade de obtenção de conhecimento pelo ser humano. Embora só nós humanos entendamos a semântica naturalmente, não somos capazes de processar terabytes de conteúdo de forma rápida para obter conclusões, fazer deduções e inferências do conteúdo presente na Web. A Web semântica é um grande auxílio cognitivo ao ser humano e com certeza um mecanismo de compartilhamento e construção de conhecimento muito mais poderoso que a Web baseada apenas em documentos e links para outros documentos.

Dado esse contexto, [5, 6] apresenta um *framework* semântico com *workflow* flexível projetado para facilitar a produção de conteúdo em uma redação jornalística. O *framework* possui três dimensões tangíveis que se relacionam sendo elas uma dimensão de conhecimento, outra de conteúdo e outra de produção de notícias conforme a Figura 1.1 .

O trabalho aqui proposto tem como objetivo ser a implementação inicial da dimensão de conteúdo desse workflow, ou seja, a construção de um protótipo de um sistema de gerenciamento de conteúdo aplicável no contexto jornalístico que gere anotações semânticas para serem utilizadas dentro do próprio sistema e também para serem compartilhadas na Web semântica de forma que um jornalista possa escrever, anotar, buscar e publicar artigos com apoio do processamento de informações semânticas. Essas anotações semânticas são feitas baseadas em uma ontologia de domínio exposta em [7] e que foi escolhida por ser produto de pesquisa do mesmo grupo de pessoas envolvidas no desenvolvimento do framework exposto em [5, 6] que é o contexto no qual esse trabalho de conclusão de

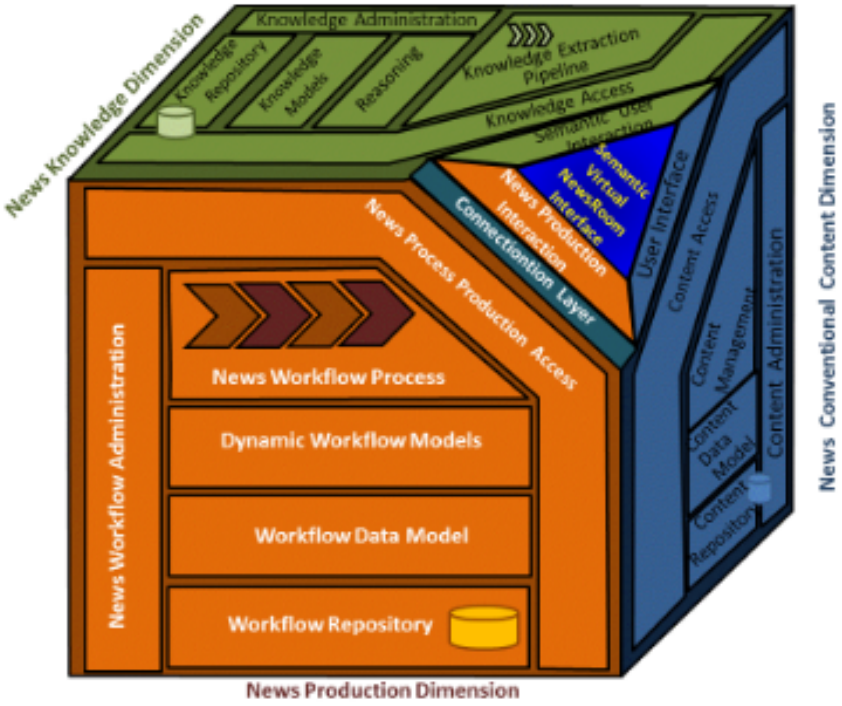

Figura 1.1: Dimensões do workflow (Fonte: [5, 6]).

curso está incluso. A pesquisa e o desenvolvimento do trabalho são feitos com base na metodologia DSR [1].

1.1 Objetivos

1.1.1 Objetivos gerais

O objetivo geral deste trabalho é implementar um protótipo funcional de um sistema de gerenciamento de conteúdo para uma redação jornalística com suporte para a construção de anotações semânticas de textos jornalísticos e também o reuso dessas anotações para auxiliar os usuários dentro do próprio sistema. Também dar suporte ao compartilhamento dos artigos e suas respectivas anotações semânticas na Web.

1.1.2 Objetivos específicos

1. Construir no protótipo um módulo que possibilite a anotação semântica semi-automática dos artigos, ou seja, gerar automaticamente, retirar ou adicionar anotações manualmente ;

2. Desenvolver e testar algoritmo de consultas para sugestão automática de textos relacionados semânticamente ordenados por grau de relacionamento nas telas de criação e edição dos artigos;

3. Desenvolver algoritmo de consultas para busca de artigos baseado nos campos de título, subtítulo, autor, editoria e nas anotações semânticas criadas por meio de URIs ou textualmente;

4. Implementar recursos HTTP que ofereçam um HTML com o artigo e um arquivo RDF de suas anotações semânticas de forma a possibilitar a publicação do artigo como uma contribuição a Web semântica;

5. Por fim ter um protótipo de sistema de gerenciamento de conteúdo com suporte semântico para criação e edição de artigos;

1.2 Estrutura do Documento

A estrutura do documento se baseia no apresentado em [8] e está organizada da seguinte maneira seguinte maneira:

- **Capítulo 1:**Expõe e contextualiza o problema. Introduz os objetivos da proposta;
- **Capítulo 2**: Apresenta a metodologia utilizada para guiar o desenvolvimento da pesquisa e do artefato
- **Capítulo 3**: Expõe de forma breve as tecnologias que dão suporte a Web semântica. Expõe o processo de revisão da literatura feita para obter o embasamento teórico necessário para o desenvolvimento desse trabalho. Também apresenta alguns trabalhos relacionados importantes para a construção de anotações semânticas a partir de segmentos de texto
- **Capítulo 4**: Descreve a arquitetura da solução, as etapas da implementação, os módulos e o funcionamento do artefato;
- **Capítulo 5**:Expõe resultados e propõe possíveis trabalhos futuros.

Capítulo 2

Revisão Bibliográfica

Este Capítulo apresenta brevemente o contexto no qual esse trabalho está inserido, versa sobre o referencial teórico necessário para a compreensão do trabalho e também expõe trabalhos que também foram desenvolvidos para anotação semântica de textos e que foram selecionados por meio de revisão da literatura. O foco das citações são os métodos de relacionamento dos textos com as ontologias e as formas de construção das anotações semânticas. Também é detalhado o que foi feito na revisão da literatura na seção . A seção 2.1.1 expõe breves definições de ontologia. A seção 2.1.2 tem uma breve introdução sobre a Web Semântica e suas tecnologias, sendo essas o padrão RDF, triplas rdf em 2.1.3, OWL em 2.1.4 e SPARQL em 2.1.5. A defição e exemplificão anotações semânticas são vistas em 2.1.8.

2.1 Conceitos Básicos

2.1.1 Ontologia

Segundo [9], o termo “ontologia” tem origem em um ramo na Filosofia, mais especificamente na Metafísica e trata da natureza e relações do "ser"e da "existência".

Na Ciência da Computação, ontologia não diz respeito ao "ser"ou a existência. Thomas Gruber em [10] define ontologia como sendo uma especificação de um vocabulário de representação para um domínio compartilhado de discurso, ou seja, definições de classes, relações, funções e outros objetos. [11] define uma ontologia como sendo uma especificação formal e explícita de uma conceitualização compartilhada.

Em [12] traz-se uma definição formal. Neste uma ontologia é uma 6-upla ”O” tal que O=(C,A^C,R,A^R,H,X), onde C é um conjunto de conceitos, A^c é uma coleção de conjuntos de atributos, R é um conjunto de relacionamentos, A^R é uma coleção de conjuntos de atributos, um para cada relacionamento de R, H é um conjunto de relações superclasse-

subclasse que representa a hierarquia dos conceitos e X é um conjunto de axiomas. Dessa forma, seja c_i um conceito em C, seu conjunto de atributos é denotado por $A^C(c_i)$. Cada relacionamento r_i<c_p, c_q> contido em R entre dois conceitos c_p e c_q também tem um conjunto de atributos denotado por $A^C(r_i)$. H é derivado de C e <c_p,c_q> pertence a H se c_p é superclasse de c_q. Cada axioma em X é uma restrição em atributos de conceitos ou entre os próprios conceitos.

É possível construir ontologias que não tenham restrições e portanto o conjunto X citado na definição anterior é vazio. Devido a isso pode-se separar as ontologias existentes em dois conjuntos sendo esses o de ontologias leves e pesadas[13]. Sendo as ontologias leves aquelas que não possuem restrições e pesadas as que possuem. Seiji Isotani em [9] define ontologias leves como sendo aquelas que não se preocupam em definir detalhadamente cada conceito representado mas apenas a relação hierárquica entre conceitos. Para ontologias pesadas, [9] diz que além de conter a hierarquia dos conceitos, elas também possuem a representação rigorosa da semântica entre os conceitos, ou seja, uma definição formal da semântica entre os conceitos e suas relações. Para criar bases de conhecimento reusáveis e compartilháveis é fundamental definir ontologias pesadas [9].

Além disso, [14] classifica ontologias em quatro tipos de acordo com níveis de generalização. Esses são ontologias de alto nível, ontologias de domínio, de tarefas e de aplicação.

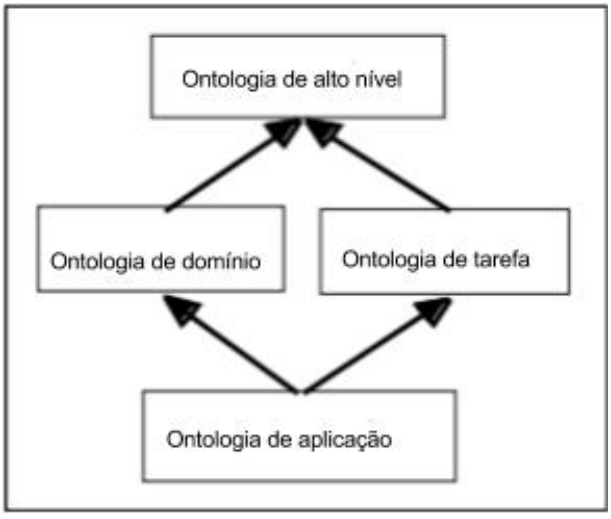

Figura 2.1: Tipos de ontologia (Fonte: [14]).

Ontologias de alto nível descrevem conceitos gerais que são independentes de domínios e problemas particulares como espaço tempo, matéria, eventos e ações.

Ontologias de domínio descrevem o vocabulário relacionado a um domínio de conhecimento. Já as de tarefas descrevem tarefas e atividades como diagnóstico de doenças ou construir uma máquina, por exemplo.

Ontologias de aplicação descrevem conceitos que dependem de um domínio particular e de tarefas. Ou seja, descrevem as entidades e suas atividades dentro de um domínio.

Um exemplo simples de ontologia é o contido em [15]. Que é uma ontologia de domínio para pizzas.

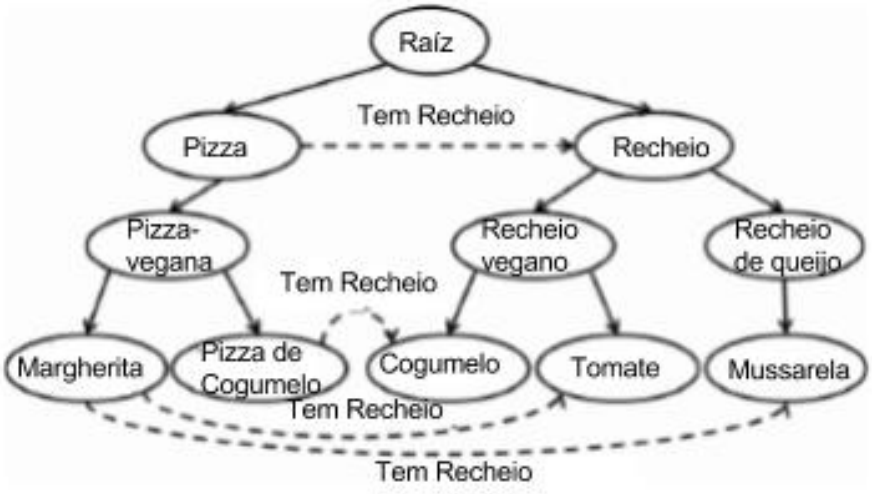

Figura 2.2: Ontologia da pizza (Fonte: [16]).

Na Figura 2.2 os círculos representam as classes, as setas contínuas representam os relacionamentos superclasse-subclasse e as setas pontilhadas representam as relações entre as classes. Essa ontologia possui formato de árvore. Pode-se observar que pizza vegana é uma subclasse de pizza e que as pizzas de cogumelo e margherita são vegetarianas. As setas tracejadas também permitem concluir que a marguerita tem os recheios tomate e queijo mozzarela, assim como concluir que a pizza de cogumelo possui recheio de cogumelo. As heranças ocorrem de maneira similar nos recheios. É trivial observar na ontologia que cogumelo e tomate são recheios vegetarianos e mussarela um recheio de queijo.

2.1.2 Web Semântica

O termo Web Semântica foi apresentado pela primeira vez em [4] em 2001 como sendo uma nova forma de conteúdo Web que possui significado para computadores.

Tim Berners-Lee em [4] também expõe os principais conceitos relacionados a Web Semântica que são:

- expressão de significado;

- representação de conhecimento;
- ontologias;
- agentes.

A expressão de significado diz respeito a construção de estruturas para o conteúdo com significado das páginas Web. Estas estruturas criam um ambiente onde softwares possam realizar tarefas que necessitem de compreensão semântica não ficando apenas no nível de comparação de palavras chave. Um exemplo de utilização dessas estruturas semânticas seria uma máquina "interpretar"que a palavra mangueira em uma página Web diz respeito a escola de samba mangueira e não a planta frutífera e muito menos ao objeto que usamos para molhar coisas. Além disso, "saber"que um texto que fala de mangueira está relacionado ao carnaval do Rio de Janeiro e a cultura brasileira.

A Representação do conhecimento diz respeito as tecnologias que são usadas para representar as estruturas semânticas citadas anteriormente. Dessas tecnologias as mais importantes são Resource Description Framework (RDF) que é um padrão para criar correlações entre dados na Web e será melhor explicado na seção 2.1.3 , o Extensible Markup Language (XML) que é uma linguagem de marcação assim como o HTML e o Universal Resource Identifier (URI) que é um padrão para construção de identificadores únicos para conteúdos da Web.

As ontologias, ainda em [4] são apresentadas como a solução para o problema da construção de inferências sobre o conteúdo da Web como, por exemplo, concluir que dois termos diferentes tem o mesmo significado pois seriam associados ao mesmo conceito em uma ontologia ou inferências mais complexas como se uma cidade referenciada na web semântica fica nos Estados Unidos, então o formato para a inserção do endereço postal de uma casa nessa cidade por um usuário na Web deve seguir o padrão norte americano.

Os agentes da Web semântica ainda segundo [4] seriam programas capazes de processar o conteúdo semântico e compartilhar o resultado com outros programas da rede. [17] cita um exemplo de um agente inteligente que realizaria tarefas de um agente de viagens. Baseado no destino, data de chegada e partida, o agente procuraria recursos na internet que estão relacionados àquele destino e que são lugares ou eventos, além de acessar informações de temporalidade, como horário de abertura e fechamento dos locais, data de realização de eventos e previsões do tempo. Por fim, o agente dá sugestões de roteiros completos para a viagem, vôos e outros transportes e até informa sobre pessoas conhecidas que estão próximas ao destino baseado em redes sociais. Esse mesmo agente pode publicar informações sobre a viagem e costumes do viajante para que outros agentes direcionem propagandas e sugestões ao mesmo.

Em [18] traz-se uma explicação simples sobre a motivação do surgimento da Web Semântica. Na web atual os documentos são compartilhados entre aplicações, mas os dados estão em posse de aplicações que mantém esses dados para si e de forma compreensível apenas para si mesmas. A Web semântica é a evolução da Web de documentos para a Web de dados e então os dados seriam relacionados entre si usando a mesma estrutura que a Web de hoje utiliza para relacionar documentos. [18] também traz uma lista de perguntas e respostas simples e completa sobre a Web semântica.

2.1.3 Triplas RDF

Resource Description Framework (RDF)

Baseado em [19] pode-se dizer que o Resource Description Framework (RDF) é uma linguagem declarativa e um padrão de utilização do XML para representar sentenças sobre propriedades e relacionamentos entre recursos da Web de maneira legível por humanos e também processável por máquinas. Esses recursos podem ser qualquer objeto (texto, figura, vídeo e etc), desde que possuam um URI. Os relacionamentos citados anteriormente são construídos por meio de triplas conhecidas como triplas RDF e melhor explicadas adiante.

Em [18] tem-se que o RDF é um padrão de relacionamento de dados na Web. Seus objetivos são facilitar a construção de relacionamentos entre dados mesmo com dados baseados em esquemas diferentes e suportar a evolução dos esquemas sem que nenhuma mudança seja feita nos agentes que consomem os dados.

Em [20] tem-se um bom material para aprofundamento em RDF.

Ainda em [18] temos que o RDF estende a estruturas de *links* da internet de forma a ser possível usar URI para nomear relacionamentos entre dois *links* e essas são as triplas RDF. Esse simples modelo permite que dados estruturados e semi-estruturados sejam misturados, expostos e compartilhados por aplicações diferentes.

Essa estrutura forma um grafo orientado e rotulado onde as arestas são predicados e os vértices são sujeitos e objetos assim como na Figura 2.4.O arquivo da ontologia da Figura 2.4 está em `https://www.dropbox.com/s/lmgiytwhrtjlg3k/root-ontology.owl?dl=0` e pode ser melhor vizualizado utilizando o programa protegé [21].

A tripla <sujeito><predicado><objeto>(Figura 2.3) é uma tripla RDF.

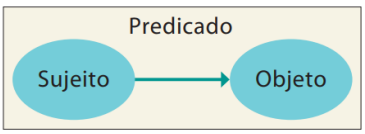

Figura 2.3: Tripla RDF (Fonte: [9]).

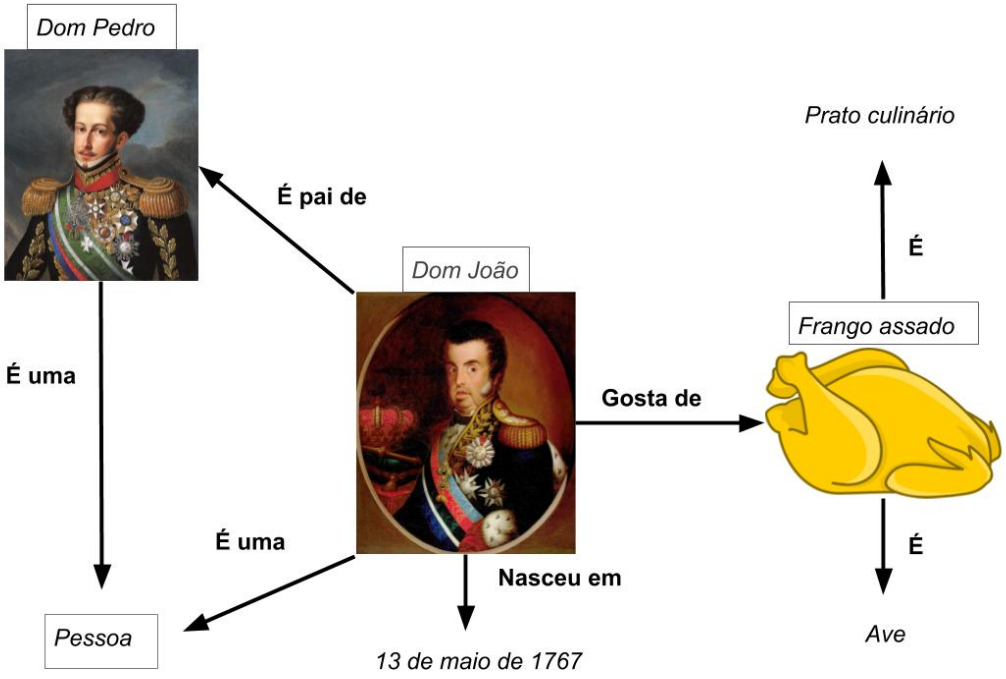

Figura 2.4: Grafo RDF. Fonte: Autor.

2.1.4 OWL

Em [22] tem-se que a Web Ontology Language (OWL) é uma linguagem para representar ontologias na Web, ou seja, representar conhecimento rico e complexo sobre coisas, grupos de coisas e relacionamentos entre coisas de forma legível por humanos e processável por máquinas. Cada ontologia representada em OWL pode ser publicada na web e pode referenciar e ser referenciada por outras ontologias.

A OWL tem maior poder para dar suporte ao processamento automatizado do conteúdo da Web em comparação com as tecnologias XML e RDF, tendo em vista a sua capacidade de representar conteúdo semântico de forma não ambígua [23].

2.1.5 SPARQL

2.1.5 Assim como o *Structured Query Language (SQL)* o *Protocol and RDF Query Language (SPARQL)* também é uma linguagem de consulta. Entretanto o SPARQL é uma linguagem de consulta para o RDF, ou seja, para consultas em triplas RDF. Em [24] tem-se que o SPARQL pode ser usado para construir consultas sob fontes de dados diversas quer essas fontes estejam diretamente em RDF ou que possam ser enxergadas

como RDF através de um middleware mas que não estejam de fato em RDF. Os resultados de consultas SPARQL podem ser conjuntos ou grafos RDF [24].

A pergunta de qual é a diferença entre o SPARQL e o SQL remete a pergunta de qual é a diferença entre um banco relacional e o RDF [25]. O RDF pode ser usado para construir um banco relacional e nesse caso o SPARQL teria a mesma aplicação do SQL [25]. Entretanto o RDF foi desenvolvido para construir grafos.

[26] contém um exemplo sobre consultas com as mesmas restrições construídas em SQL e SPARQL:

2.1.6 Exemplo de consulta SQL

```
SELECT UNIQUE E.SALARY
FROM EMPLOYEES AS E JOIN DEPARTMENTS AS D
WHERE E.ID=D.MANAGER;
```

2.1.7 Exemplo de consulta SPARQL

```
SELECT ?salary
WHERE
?e rdf:employees/column#salary ?salary.
?d rdf:departments/column#manager ?e.
```

Nas duas consultas teremos um conjunto contendo o salário do empregado que é o gerente. Entretanto enquanto na consulta SQL compara-se colunas de duas tabelas diferentes (E.ID=D.MANAGER), no SPARQL compara-se o recurso que está sendo apontado por dois URI's, ou seja, procura-se o recurso rdf:employees que também é apontado por rdf:departments/column#manager em um grafo. [26] é um estudo detalhado sobre as diferenças entre as duas tecnologias.

2.1.8 Anotações semânticas

O termo "anotação"implica, genericamente falando, em atrelar dados a outras porções de dados[27]. No dicionário da língua portuguesa, uma anotação é uma indicação escrita breve, apontamento, nota ou chamada. Ou também uma série de comentários sobre produção literária, artística ou científica, ou seja, de fato uma porção de dados que dizem respeito a outros dados. [27] também contém definições formais sobre anotações semânticas. A mais geral delas diz que uma anotação A é uma 4-upla(a_s,a_p,a_o,a_c), onde a_s é o

sujeito da anotação, ou seja, o dado que está sendo anotado, a_p é o predicado, ou seja, o tipo da relação entre a_s e a_o, a_o é a entidade que se relaciona a a_s pelo predicado a_p, e a_c é o contexto na qual a anotação está inserida. O predicado, o sujeito e o contexto podem ser tanto formais quanto informais. Anotações de rodapé em um livro são consideradas anotações semânticas. Um autor pode anotar na versão original de um livro o significado de uma palavra e atribuir um contexto como data e local no qual aquela anotação foi criada ou tem sentido. Já uma anotação completamente formal seria uma na qual todos os itens dessa 4-upla sejam um Universal Resource Identifier (URI).

```
<rdf:Description rdf:about="http://example.org/bob#me">
   <rdf:type rdf:resource="http:xmlns.com/foaf/0.1/Person"/>
   <schema:birthDate rdf:datatype="http:www.w3.org/2001/
    XMLSchema#date">1990-07-04<schema:birthDate>
   <foaf:knows rdf:resource="http://example.org/alice#me"/>
   <foaf:topic_interest
    rdf:resource="http://www/wikidata.org/entity/Q12418">
</rdf:Description>
```

Figura 2.5: Exemplo de anotação em RDF (Fonte: [9]).

A Figura 2.5 mostra um trecho de anotação formal baseada em ontologia feito em RDF. Pode-se observar que o código descreve o recurso Bob, identificado pelo URI <rdf:Description rdf:about=”http://example.org/bob#me” >, na linha seguinte tem-se o tipo de recurso descrito que referencia um conceito da ontologia FOAF presente em [28], em seguida é dada a data de nascimento de Bob na linha <schema:birthDate rdf:datatype=”http:www.w3.org/2001/XMLSchema#date” >1990-07-04<schema:birthDate> que está no formato descrito em ”http:www.w3.org/2001/XMLSchema#date”. Na linha seguinte é dito que bob conhece alice e em seguida que tem interesse pelo tópico descrito pelo URI ”http://www/wikidata.org/entity/Q12418”. Nessa anotação é possível identificar várias triplas <sujeito><predicado><objeto> que são:
<bob><é><pessoa>
<bob><nasceu em><1990-07-04>
<bob><conhece><alice>
<bob><tem interesse em><”http://www/wikidata.org/entity/Q12418” >

2.2 Trabalhos relacionados

Tendo como base que o foco desse trabalho é a construção de anotações semânticas de segmentos de texto, essa seção se restringe a apresentar artigos científicos que também têm objetivos próximos a esse além de expor a revisão da literatura que foi feita para escolher esses artigos científicos. A anotação semântica de segmentos de texto possui três fases não facultativas que são a identificação de conceitos em um texto, o cálculo ou a dedução de uma similaridade semântica entre um conceito do texto e um outro referencial, como por exemplo um conceito presente em uma ontologia e a construção da anotação em si. Visto isso, o foco aqui é expor de forma resumida trabalhos com formas relevantes de identificar conceitos em texto ou de descobrir similaridade semântica ou de construir anotações semânticas.

2.2.1 A pesquisa

As buscas foram feitas sobre artigos científicos em inglês disponíveis no Google Acadêmico [29] ou na rede CAPES.

A abordagem inicial foi a de usar uma chave de busca abrangente e depois ir restringindo-a a partir dos resultados não relacionados ao foco dessa pesquisa que fossem encontrados.

O objetivo da pesquisa foi: Encontrar formas de processar textos e anota-los semanticamente de forma semi-automática, ou seja, auxiliada por computador mas feita por humanos. Além disso essas anotações precisariam ter como referêncial uma ontologia. Portanto a chave de busca mais abrangente inicial foi *ontology annotation*, que traria artigos com os termos *ontology* e *annotation*. Após isso foi feita uma leitura nos títulos dos resultados para identificar palavras que retirassem os resultados não congruentes com a pesquisa, por exemplo, artigos sobre filosofia que contivessem o termo "ontology". Após a revisão em várias páginas da busca foi elaborada a seguinte chave de busca:
ontology annotation -photo -video -geographic -ethics -morality - "ontology generation" -wikipedia -philosophy

Vários artigos se referiam a filosofia ou sobre a anotação de fotos e vídeos. O sinal "-"(menos) indica a retirada da palavra subsequente. Essa última chave não é boa pois não discrimina o conteúdo do título do conteúdo do texto. Algum texto que tivesse a palavra "video" e falasse de anotação semântica de textos não apareceria, por exemplo. Após a observação dos resultados da pesquisa e manipulação da chave de busca para reduzir os resultados indesejados chegou-se na seguinte chave:
"semi-automatic"
intitle:ontology|ontologies
intitle:annotation|annotating|annotations intext:text

intext:semantic -intitle:genetic -intitle:gene
- intitle:photo -intitle:video -intitle:image|images
- intitle:paintings -intitle:music -intitle:metaontology

Essa chave trás os artigos que tenham o termo ”semi-automatic” exatamente em algum lugar e filtra melhor artigos que tratam de anotação automática ou manual. O trecho intitle:ontology|ontologies retorna artigos que tenham alguma das palavras ontology ou ontologies no título assim como no trecho intitle:annotation|annotating|annotations. O trecho intext:semantic retorna artigos que tenham a palavra semantic no texto. A retirada de palavras foi feita apenas no título a fim de evitar a perca de conteúdo importante devido a variedade de palavras que é possível encontrar no texto. Além disso, retirar artigos a partir de palavras do texto pode acabar excluindo resultados interessantes visto que dentre as várias palavras de um artigo a palavra ”metaontology” poderia ser citada mesmo que o artigo tratasse de anotação de textos e não de ontologias como sugere o termo ”metaontology”. A partir desse ponto ficou incerto retirar palavras sem a possibilidade de perder bons resultados. A chave ainda trazia por volta de 600 resultados. Os mesmos foram filtrados lendo-se o título e quando necessário o resumo dos artigos. Por fim restou uma lista de 128 artigos que tiveram seu conteúdo avaliado totalmente ou em parte de forma orientada aos objetivos da pesquisa.

2.2.2 Sistema semântico de gerenciamento de conteúdo

Um sistema de gerenciamento de conteúdo semântico (CMS) está relacionado a uma dimensão de criação de conteúdo que não necessáriamente é gerenciada completamente dentro do mesmo. Além disso, em um sistema que está conectado a Web semântica e portanto processa e referencia conteúdo semântico processável por máquinas, existe uma dimensão de gestão do conhecimento que dá suporte a esse processamento semântico e está representado com as tecnologias da Web semântica. Essa dimensão também não é completamente abarcada pelo CMS. [5] explora essas dimensões no contexto de um framework semântico que dá suporte a uma redação jornalistica.

O trabalho apresentado aqui é uma proposta para a dimensão de conteúdo do *framework* apresentado em [5]. Suas dimensões são mostradas como as faces de um cubo na Figura 1.1 sendo uma dimensão de conteúdo, outra dimensão de processo de produção de notícias e uma de gerenciamento de conhecimento que provê uma interface de busca para as anotações semânticas feitas nas outras duas dimensões possibilitando a melhoria e criação de processos de produção de notícias. Na dimensão de conteúdo, o *framework* proposto em [5] dá suporte a elaboração de artigos e documentos. Além disso os jornalistas podem anotar semanticamente os textos, obter trabalhos semanticamente relacionados e referenciar os textos relacionados dentro do próprio texto. É também explicitado que

a dimensão de conteúdo é implementada por um CMS. As dimensõe de conhecimento e produção de noticias são melhor detalhadas em [5].

2.2.3 Anotação semântica

Dentre os resultados obtidos,percebe-se uma interessante abordagem para à construção de esquemas para registros de anotações em [30]. Esse explica formas de denotar a relação de segmentos de textos a conceitos de ontologias utilizando Xpointer[31], que é um sistema para endereçamento de elementos de conteúdo XML na internet.

Um bom registro de anotação possui metadados também relacionados a ontologias e isso é muito bem explorado nesse artigo. Em suma, o resultado do trabalho é uma ontologia para anotações semânticas chamada de AO(Annotation Ontology)[32]. Nos registros de anotação semântica, além da ontologia AO também são usadas outras como FOAF[28] e PAV[33].

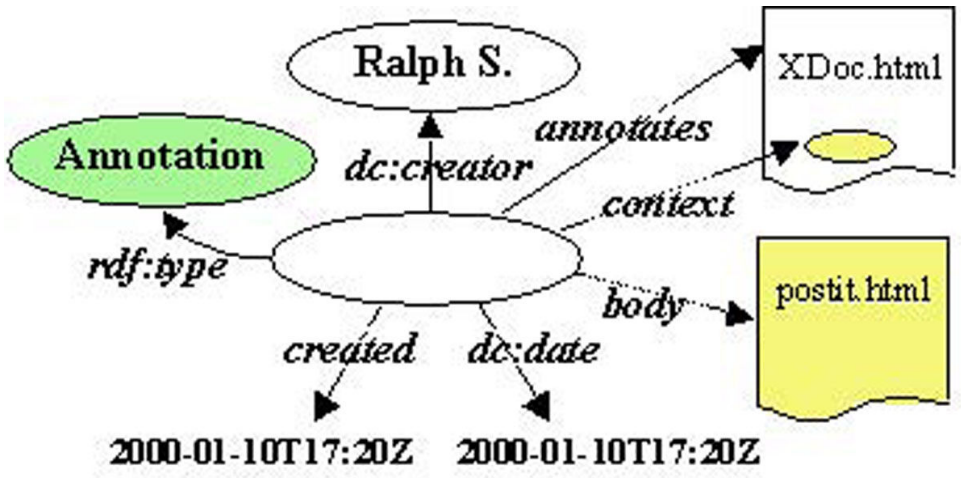

Figura 2.6: Registro de anotação (Fonte: [30]).

A Figura 2.6 é um dos modelos de registro de anotação propostos em [30] e também o mais próximo do modelo adotado nesta monografia. O círculo central é o sujeito da anotação que no caso é o artigo que está sendo anotado. Os predicados estão sobre as linhas e os objetos apontados pelas setas.

A Figura 2.7 é a representação em rdf do modelo apresentado com as respectivas ontologias representadas na forma <ontologia ou namespace:predicado:recurso = url do resurso>. Onde recurso é o sujeito da tripla sujeito, predicado e objeto. Observando essa imagem podemos identificar que o trecho de texto referenciado por
serv1example.com/some/page.html# xpointer(id(”Main”))/p[2]
referencia o tópico presente em purl.obolibrary.org/obo/PRO_000004615 e a anotação

```
< rdf : Description rdf : about = "http : / / my.example.org / ann / 29332" >
        < rdf : typerdf : resource = "& aot; Qualifier " / >
        < rdf : typerdf : resource = "& ao; Annotation" / >
        < rdf : typerdf : resource = "& ann; Annotation" / >

< ann : context > http : / / serv1.example.com / some / page.html # xpointer(i
d("Main") / p[2]) < / ann : context >
        < aof : annotatesDocument
rdf : resource = "http : / / tinyurl.com / ykjn87p" / >
        < ao : hasTopic
rdf : resource = "http : / / purl.obolibrary.org / obo / PRO _ 000004615" / >
        < pav : createdOn > 2010 - 03 - 21 < / pav : createdOn >
        < pav : createdBy
rdf : resource = "http : / / www.hcklab.org / foaf.rdf # me" / >
< / rdf : Description >
```

Figura 2.7: Registro de anotação em formato RDF (Fonte: [30]).

em questão foi criada pela entidade representada em www.hacklab.or/foafrdf#me. Além disso esse registro de anotação também contém referências semânticas sobre o que é ele próprio, ou seja, links na Web semântica que o descrevem. As mesmas estão na primeira *tag* RDF e nas seguintes indentadas a ela.

2.2.4 Desambiguação terminológica

Existem problemas naturais ao utilizar ontologias de domínio para anotar segmentos de texto. [34] cita alguns desses problemas. Dentre eles é importante salientar os seguintes:

1. A terminologia difere em locais diferentes. Por vezes até na mesma língua
2. Termos da ontologia podem não ter tradução em outros idiomas
3. Um segmento de texto pode ter mais de um significado na linguagem ou mesmo dentro do próprio domínio. Sendo necessário fazer um desambiguamento antes da anotação
4. Alguns conceitos possuem outros conceitos embutidos dentro de si. Um exemplo seria *carro de boi* que contém os conceitos carro e boi contidos em sí mas que juntos formam outro conceito.

Neste também é apresentada uma abordagem para a construção de anotações semânticas de segmentos de texto baseada em uma ontologia de domínio. A arquitetura da solução

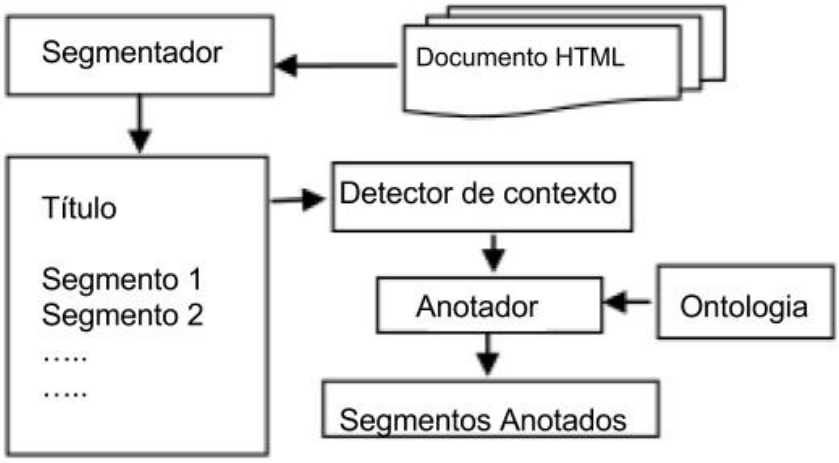

Figura 2.8: Diagrama do processo de anotação (Fonte: [35]).

deixa claro os passos executados na busca pelos termos, seu desambiguamento e anotação conforme a Figura 2.8.

O documento é segmentado em partes mínimas que façam sentido na linguagem. No português, por exemplo é interessante segmentar períodos e retirar *stopwords* [36], que em suma são palavras que não causam perca de informação ou significado quando retiradas de um segmento de texto. Os segmentos são enviados a um detector de contexto, que no caso em questão será identificado por algumas palavras encontradas partindo-se do pressuposto que essas palavras determinam um contexto dentro do domínio da ontologia. O "anotador" irá procurar por casamentos de segmentos do texto com os conceitos da ontologia e seus sinônimos que estão previamente registrados em um *thesaurus*, que é um dicionário de sinônimos e conceitos relacionados. A saída do algoritmo é um conjunto de segmentos anotados.

Uma tarefa difícil, mas importante para se alcançar um resultado muito melhor é levar em consideração os sinônimos dos conceitos encontrados no texto. [37] ressalta essa importância com o fim de aumentar o grau de detecção de similaridades semânticas tanto na anotação quando em buscas semânticas que sejam feitas usando as anotações.

Também ressalta a importância da etapa de detecção de contextos ou desambiguamento, que assim como nesta monografia são feitos por um ser humano. Um exemplo desse desambiguamento é o que deve ser feito do trecho "A zika é muito perigosa", pois nesse caso "Zika" diz respeito a doença Zika e não ao Zika Vírus ou a outro sentido como o de maldição ou baixo astral que existem na linguagem coloquial brasileira e por isso não deve ser associada aos mesmos.

De forma parecida a abordagem proposta nessa monografia, também é feita uma comparação textual de conceitos e sinônimos da ontologia com segmentos do texto.

É claro que além de anotar os conceitos do texto, é possível ir além e anotar inferências que se faz do texto ou dos próprios conceitos em si. Em [35], além de anotar conceitos do texto que tenham correspondência com uma ontologia de domínio, são geradas novas anotações a partir de uma técnica de raciocínio que obtém conceitos relacionados que podem ser inferidos. As anotações criadas a partir de inferências são associadas a meta-anotações que tem o objetivo de "explicar" como as mesmas foram inferidas.

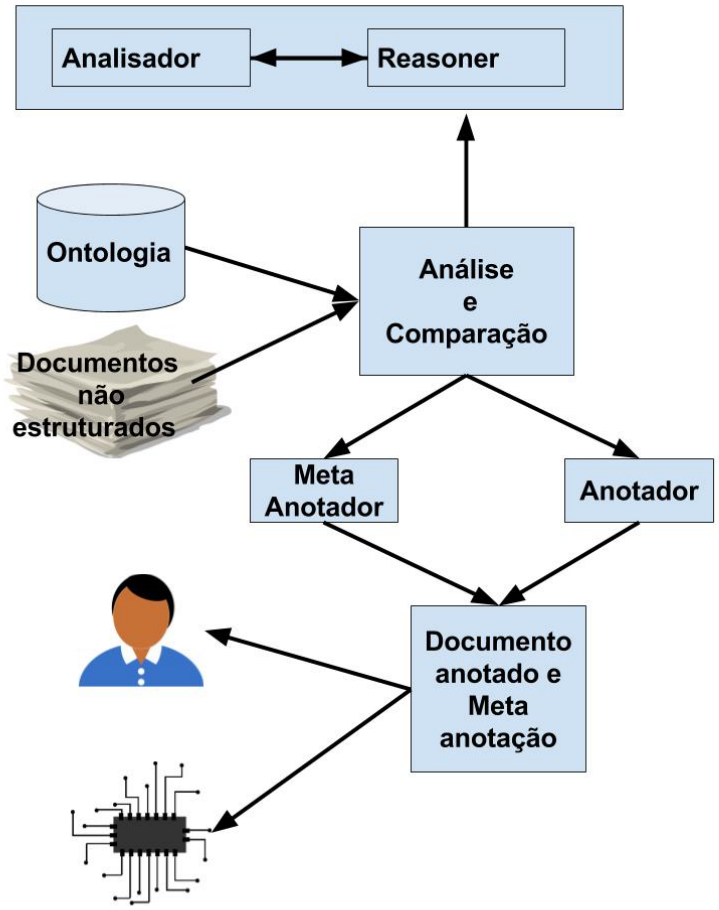

Figura 2.9: Diagrama do processo de anotação com *Reasoner* (Fonte: [34]).

Um exemplo simples de inferência seria anotar que José é pai de Maria. Se temos na ontologia que José é mesmo que Zé. Então pode-se inferir que Zé é pai de Maria.

O resultado da anotação pode ser aproveitado por humanos ou máquinas, como explicitado na Figura 2.9.

Os dicionários de línguas sempre foram um referencial para humanos. Dicionários de línguas processáveis por máquinas têm se tornado realidade. Um exemplo é a [38], que é uma base de dados léxica processável por máquinas. [39] traz uma abordagem diferente para construir a anotação. Para identificar de forma única os conceitos encontrados nos textos não é usada uma ontologia mas sim a Wordnet[38]. Na Wordnet, sinônimos são agrupados em conjuntos. Cada conjunto representa um conceito distinto. Os conjuntos estão interligados por relações semânticas e léxicas. A partir dos conceitos da base de dados pode-se gerar um URI's, ou seja, um identificador único na internet, para cada palavra encontrada em um texto que exista na Wordnet.

Para realizar o desambiguamento e a descoberta de contexto das palavras encontradas nos textos é usado o algoritmo WSD[40](”Word Sense Disambiguation”).

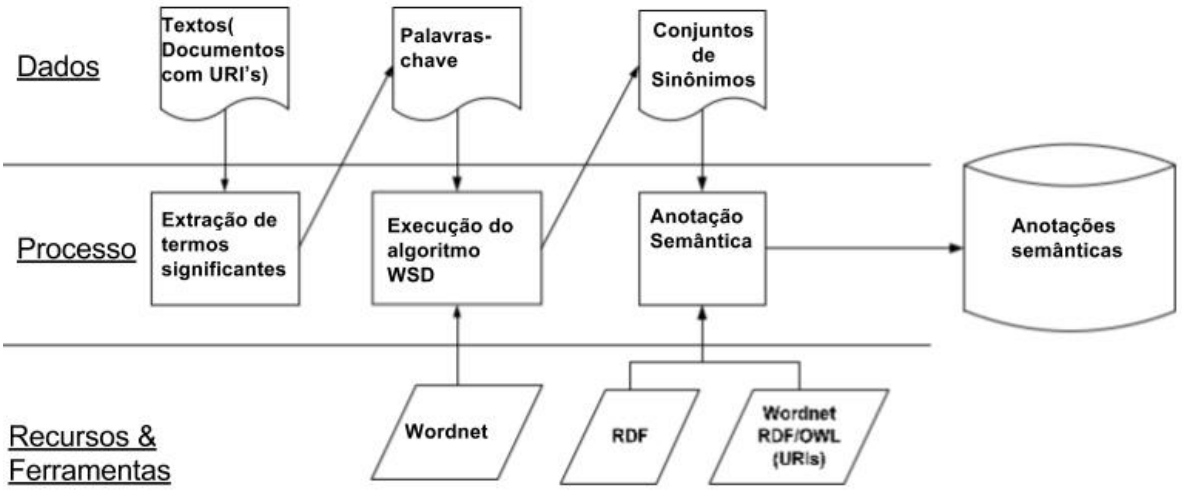

Figura 2.10: Processo de anotação semântica com uso do WSD e Wordnet (Fonte: [39]).

Na Figura 2.10, a camada de dados é relacionada aos dados da Wordnet. A de processos as tarefas de extração de conceitos do texto, relacionamento de conceitos da Wordnet com segmentos encontrados no texto e construção da anotação semântica. A camada inferior está relacionada com as ferramentas utilizadas para obtenção dos URI's dos termos encontrados nos textos e para a construção das anotações semânticas. Desse modo, dado um texto que possua um identificador, ou seja, um URI, o texto é processado por um módulo que extrai as sentenças importantes do texto usando alguma heurística como separar as sentenças em períodos ou retirar stopwords [36]. Em seguida as sentenças extraídas são processadas pelo algoritmo WSD[40](”Word Sense Disambiguation”) e os conceitos encontrados na Wordnet vão para um módulo que constrói as anotações semânticas baseado no URI do texto e nos URI's dos conceitos encontrados na Wordnet.

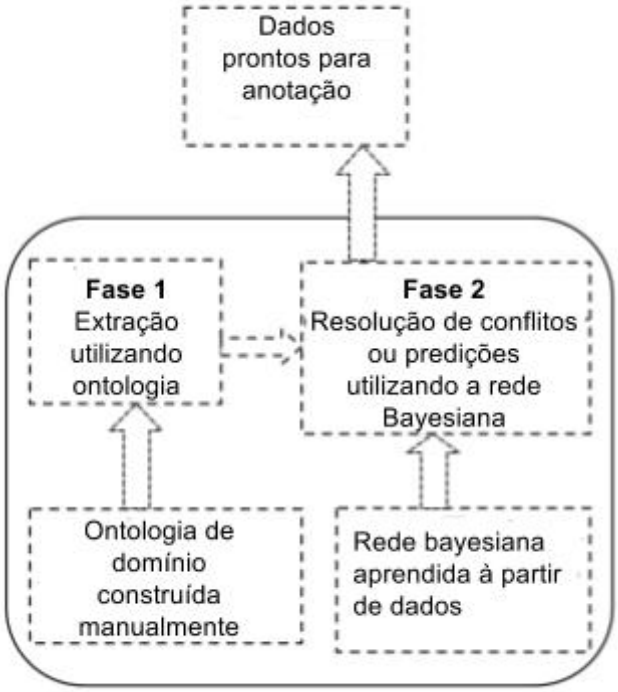

Figura 2.11: Processo de anotação semântica com uso de redes Bayesianas (Fonte: [41]).

2.2.5 Processamento automático de textos

Na área de inteligência artificial e processamento de linguagem natural existem muitas abordagens para identificar similaridades semânticas entre segmentos de texto e conceitos. O framework BNOSA[41] realiza anotações semânticas também com palavras chave contidas em ontologias de domínio mas acrescenta o uso de redes bayesianas [42]. No caso de mais de um valor ser atribuído a um conceito a ser anotado ou se um conceito não foi encontrado na ontologia, a rede bayesiana se encarrega de encontrar o valor mais apropriado para o dado conceito. Por exemplo, se na frase *manga de camisa* foram atribuídos os conceitos referentes a fruta manga e a manga de camisa, a rede bayesiana se encarrega de retirar a ambiguidade e atribuir o conceito adequado.

Em suma, dado uma ontologia e uma rede bayesiana construídas para um determinado domínio, o BNOSA provê anotação de conteúdo com bons resultados. A Figura 2.11 ilustra o funcionamento.

Na fase 1 é necessária uma ontologia de domínio construída por especialistas no domínio para que seja feita a extração dos conceitos baseado nos conceitos da ontologia. Na fase 2, uma rede bayesiana previamente construída com aprendizado de máquina [40] resolve os conflitos existentes. A saída são conceitos prontos para serem anotados.

2.3 Conclusão

Neste capítulo abordou-se os conceitos básicos e definições necessários para a compreensão do trabalho e também das referências escolhidas para o mesmo.Também apresentou trabalhos relacionados que são importantes e (ou) que influenciaram decisões tomadas para alcançar os objetivos propostos.

Para este trabalho de conclusão de curso não optou-se por opções não determinísticas como as apresentadas em [41] ou em [39] para construção de anotações totalmente automáticas. A solução apresentada aqui teve alguma inspiração em [35], entretanto o detector de contexto seria a pessoa que está anotando, ou seja, após o algoritmo de anotação sugerir anotações apenas baseado em comparação textual com os conceitos da ontologia, o ser humano especialista no domínio que estiver anotando decide quais são válidas e quais estão faltando. Já[30] foi sem dúvida o trabalho mais importante para a compreensão de como construir registros de anotação usando RDF e também o que mais ofereceu exemplos de artefatos para reúso(ontologias como Annotation Ontology [32] e FOAF[28]) que auxiliaram na elaboração da arquitetura da interface RDF para as anotações semânticas criadas no sistema.

Capítulo 3

Metodologia

Este Capítulo descreve o método utilizado na implementação do trabalho aqui proposto. O método de pesquisa utilizado foi o Design Science Research (DSR) [1]. A DSR segundo [1] é a ciência que procura consolidar conhecimento sobre um projeto e desenvolvimento de soluções para melhorar sistemas existentes, resolver problemas e criar novos artefatos, sendo artefatos coisas construídas pelo homem ou interfaces entre o ambiente interno e o ambiente externo de um determinado sistema.

3.0.1 A metodologia DSR

O método Design Science Research (DSR) define sete critérios a serem seguidos. O primeiro define que devem ser construídos artefatos viáveis, na forma de um constructo modelo, método ou de uma instanciação. O segundo trata da relevância do problema e ressalta que é importante resolver problemas importantes e relevantes para as organizações. O terceiro trata da avaliação do design e afirma que a utilidade, a qualidade e a eficácia do artefato devem ser rigorosamente demonstradas por meio de métodos de avaliação bem executados. O quarto diz respeito às contribuições da pesquisa e afirma que qualquer pesquisa conduzida pelo método da Design Science Research (DSR) deve prover contribuições claras e verificáveis nas áreas específicas dos artefatos desenvolvidos e apresentar fundamentação clara em metodologias de design. A quinta diz respeito ao rigor da pesquisa no sentido que a mesma deve ser baseada em uma aplicação de métodos rigorosos, tanto na construção como na avaliação dos artefatos. O sexto critério se trata do design como um processo de pesquisa e afirma que os meios para alcançar os fins desejados devem estar disponíveis ao mesmo tempo que satisfaçam as leis que regem o ambiente em que o problema está sendo estudado. O último critério diz que as pesquisas devem ser apresentáveis tanto para público acostumado ao domínio de tecnologia da informação quanto aos que não.

A Figura 3.1 ilustra cada etapa do DSR e suas respectivas saídas de uma forma genérica, as etapas do trabalho desenvolvido ocorreram como segue mas não obrigatoriamente de forma sequencial mas cíclica e incremental:

1. Definição do problema;
2. Revisão da literatura;
3. Pesquisa de ferramentas que possibilitassem a criação das anotações baseadas em ontologia à partir de texto usando a linguagem de programação Python;
4. implementação dos módulos que permitem gerar registros de anotação , ou seja, um grafo RDF baseado em ontologias e a partir de um texto;
5. Modelagem de aplicação;
6. Implementação da aplicação;
7. Avaliação dos resultados;
8. Comunicação de Resultados;

A definição do problema se deu na reflexão sobre os ganhos que a semântica processável por máquinas pode trazer à um sistema de gerenciamento de conteúdo, como por exemplo um sistema de gerenciamento de artigos de uma redação jornalística. Tanto na produção quanto na leitura de artigos jornalísticos existem lacunas que os sistemas não conseguem resolver apenas com o processamento dos documentos em si. Um exemplo claro dessas lacunas é uma simples artigo que um jornalista faz ao escrever uma matéria sobre a cidade de Barcelona. O mesmo quer ter referências anteriores sobre a cidade, entretanto ao pesquisar na internet em sistemas que não processem conteúdo da Web Semântica ou no próprio sistema de gerenciamento de conteúdo do jornal no qual trabalha sem suporte semântico, o jornalista terá em sua pesquisa muitos ou somente resultados sobre o time de futebol Barcelona. Em um sistema que processa dados anotados semanticamente o jornalista teria os resultados desejados mais facilmente e pouparia tempo. Esse item é melhor exposto na seção 2.2.1.

A revisão da literatura foi feita com foco em estudar formas de relacionar conceitos encontrados em um texto com conceitos presentes em ontologias de domínio e construir anotações semânticas a partir disso. A construção de anotações semânticas é necessária para atingir o objetivo 1 descrito na seção 1.1.2, que é premissa para o alcance de todos os outros objetivos e por isso a pesquisa teve esse foco. As buscas foram feitas no site Google Acadêmico [29] por artigos científicos disponibilizados gratuitamente em inglês. Esse item é melhor exposto no Capítulo 3.

O protótipo é um sistema web, então era necessário escolher um framework para desenvolvimento Web gratuito que possibilitasse o alcance dos objetivos descritos na seção 1.1.2. A pesquisa de ferramentas foi direcionada a encontrar módulos implementados na linguagem de programação Python que permitissem consultar o conteúdo de ontologias e também a criação de anotações semânticas em RDF. Dado que eram necessárias tecnologias para criação e consulta em arquivos RDF e também para consultas em ontologias para alcançar os objetivos e dada a familiaridade do autor com linguagem Python e sabendo-se da existência do framework Django [43] para o Python, pesquisou-se opções em Python que possibilitassem a construção de anotações semânticas em RDF e o trato com ontologias. Para o RDF encontrou-se o RDFLib [44] e para ontologias encontrou-se o ontospy [45] e então confirmou-se a viabilidade do Django. Dentre as opções para bancos de dados relacionais disponíveis para o Django foi escolhido o MySql por afinidade.

A modelagem do modelo de persistência dos dados foi feita levando em consideração que seria necessário armazenar os dados dos artigos como título, subtítulo, texto, autores e editorias e também de suas anotações semânticas. A persistência dos dados dos artigos e suas anotações semânticas é premissa para alcançar todos os objetivos.

A implementação da aplicação se deu de forma incremental. Primeiro foi desenvolvido um módulo que dado um texto e uma ontologia, detecte os conceitos da ontologia que estão presentes no texto completando parcialmente o objetivo 1 da seção 1.1.2. Depois iniciou-se o trabalho com o Django [43], com foco em uma tela para as funcionalidades de criação e edição de artigos. O modelo da persistência de dados evoluiu gradualmente para atender as funcionalidades implementadas, ou seja, não foi concebido antes das funcionalidades. Depois que já era possível criar e editar um artigo, adicionou-se a funcionalidade de anotar e persistir as anotações de um artigo em um arquivo RDF já em um padrão compatível com a Web semântica e por fim também replicando essas anotações no banco relacional, completando assim o item 1 dos objetivos específicos. O próximo passo foi implementar a funcionalidade de publicar um artigo, ou seja, permitir ao autor do artigo gerar uma versão com aparência diferente que é armazenada para ser apresentada aos leitores e que não é mais alterada mas também permitindo que cada artigo possa ser publicado mais de uma vez cumprindo assim parte do item 4 dos objetivos específicos. O item 3 foi cumprido em seguida com a funcionalidade de busca em artigos publicados e também uma lista com *links* para edição de todos os artigos armazenados no banco. O item 2 foi atendido em seguida com duas implementações diferentes para comparação. Por fim foi implementado um recurso HTTP que oferece o arquivo de anotações semànticas em RDF de cada artigo a partir de uma URL completando assim os itens 4 e 5. A implementação é melhor exposta no Capítulo 4.

O teste e exposição dos resultados foi em suma, testar se a aplicação possibilita que

sejam feitas de forma correta as anotações semânticas semi-automáticas de textos baseadas na ontologia do Zika presente em [7]. Também comparar de forma experimental duas possibilidades para a sugestão automática de artigos relacionados baseada nas anotações semânticas dos textos. Além disso testar a busca de artigos publicados que é feita baseada nas anotações semânticas e em atributos dos artigos, como título e editoria. Toda a exposição foi feita contrastando casos de uso e seus resultados entretanto não foram produzidas avaliações quantitativas, mas sugestões para as mesmas em trabalhos futuros. Esses itens são melhor expostos no Capítulo 6.

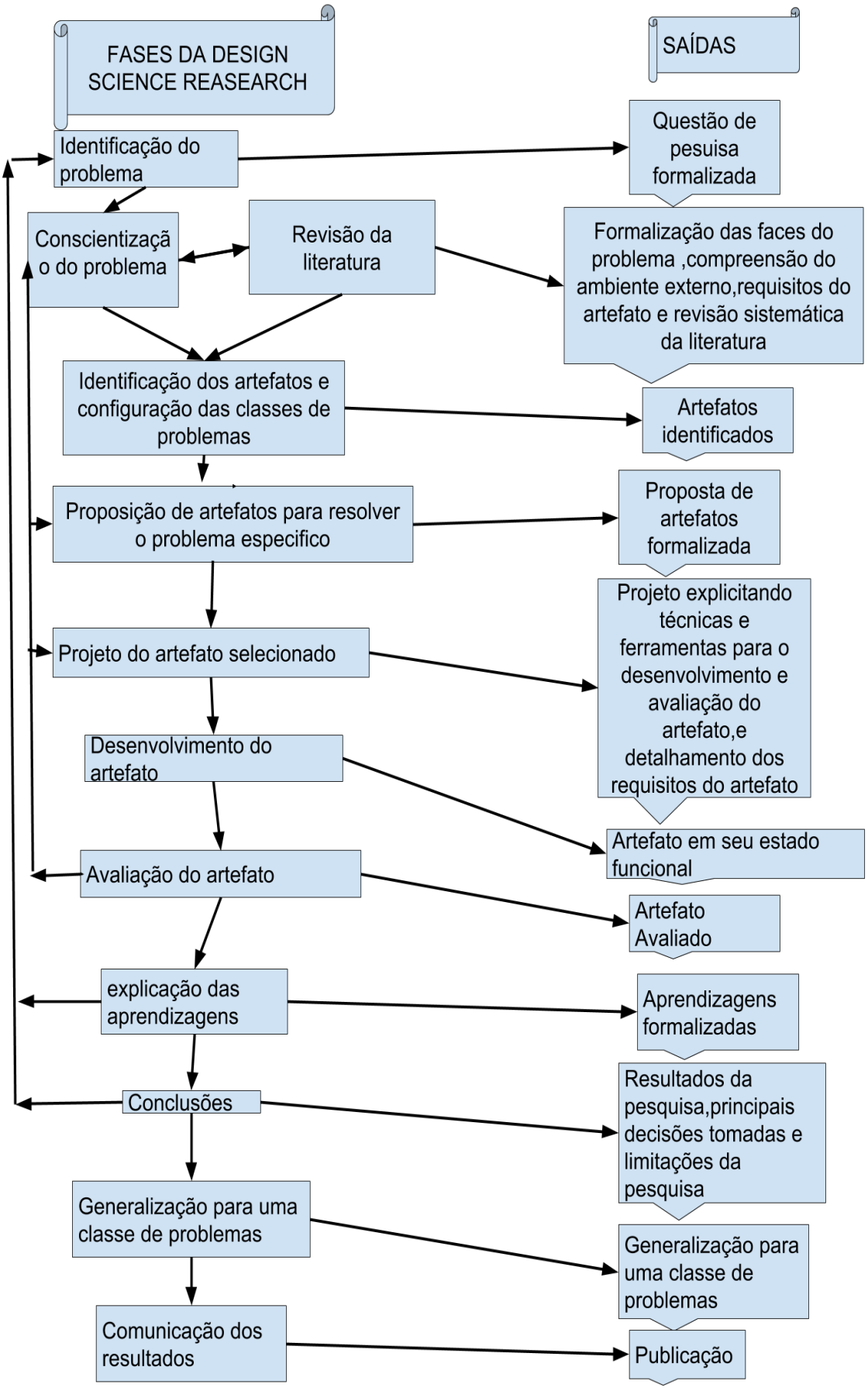

Figura 3.1: Etapas da Metodologia DSR. Adaptado de [1] .

Capítulo 4

Implementação

Este Capítulo descreve a implementação do artefato proposto nos objetivos gerais deste trabalho assim como seus casos de uso, recursos e arquitetura.

O artefato é um protótipo de sistema de gerenciamento de conteúdo desenvolvido com o objetivo de estender as capacidades de uma redação jornalística na produção de artigos jornalísticos assim como incrementar as capacidades de seus leitores no consumo das informações produzidas. As anotações semânticas semi-automáticas dos textos baseadas em ontologia produzem informações que podem posteriormente contribuir na construção de relacionamentos semânticos entre os textos que ajudam tanto na construção de novos artigos tendo como referência artigos relacionados já escritos quanto na sugestão de artigos relacionados a um dado artigo que esteja sendo lido e também nas buscas de artigos baseadas na semântica de seus textos. A ontologia utilizada como base para as anotações semânticas nessa primeira versão do artefato é a descrita no trabalho [7] e foi escolhida devido ao fato do artefato desenvolvido aqui ser uma possível alternativa prática ao que é proposto também em [7]. O artefato também pode ser apresentado como um ambiente de autoria ou produção de conhecimento baseado em ontologia e seus requisitos tem inspirações no que é apresentado em [46]. A arquitetura geral do artefato pode ser inferida da Figura 4.1. O servidor de aplicação é acessado por sistemas externos e por usuários e acessa o banco de dados e a ontologia que está armazenada localmente. Além disso referencia conceitos da Web Semântica para construir a as anotações. Alguns dados da ontologia são replicados no banco de dados e a mesma também referencia outras ontologias e conceitos na Web semântica.

4.1 Casos de Uso

A *persona* do autor tem a função de criar e revisar artigos. Essa função, além da tarefa de criar também inclui anotar editar um artigo criado. A informação produzida

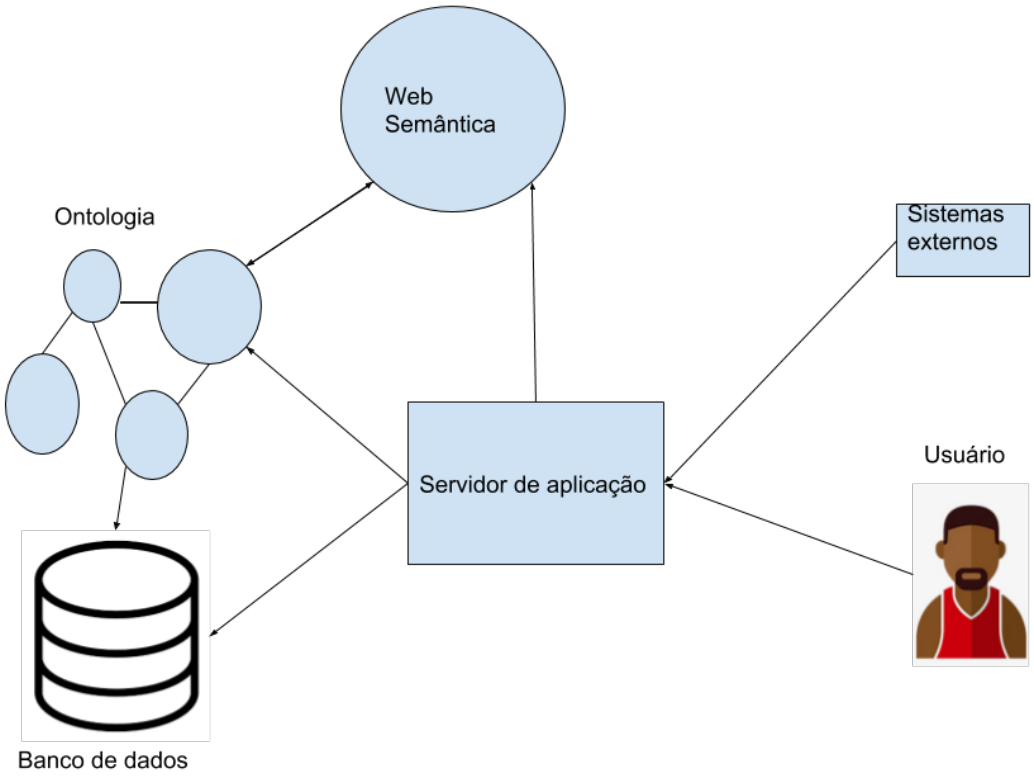

Figura 4.1: Diagrama com arquitetura geral da apliicação.(Fonte: Autor).

pelos jornalistas pode ser consumida pelo leitor por meio dos artigos publicados e uma interface de busca aos mesmos e pode ser consumida por outros sistemas semânticos por meio de arquivos RDF com as anotações semânticas de cada artigo.

A tela de criação e edição de artigos é mostrada na Figura 4.3. Os campos de título, texto e subtítulo (ou "sutian") são de inserção textual enquanto os de editoria e autores são de seleção múltipla. O botão "Anotar" executa o algoritmo de anotação sobre o texto e salva o artigo gerando uma lista de conceitos encontrados conforme a Figura 4.4. A escolha dos campos

Os conceitos podem ser desmarcados da lista por quem está anotando caso não tenham de fato uma correspondência semântica com o texto e também podem ser adicionados por meio do seletor com rótulo *Conceitos a adicionar*. Após a anotação é gerada na mesma página uma lista com os cinco artigos supostamente mais relacionados com o artigo que está sendo editado (Figura 4.13).

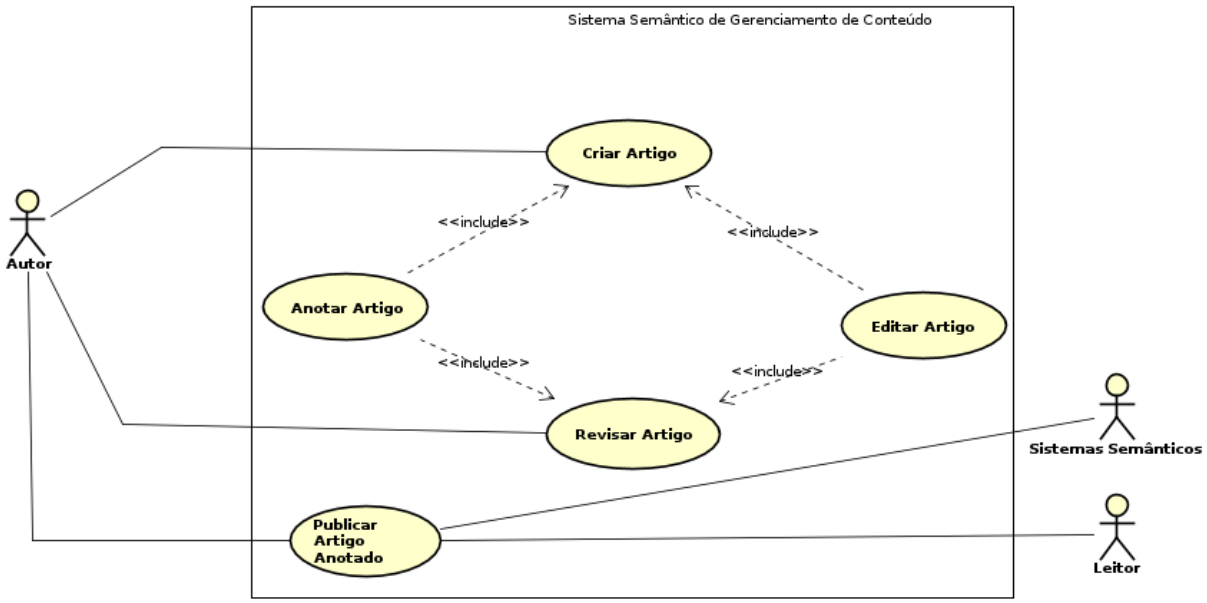

Figura 4.2: Diagrama de casos de uso.(Fonte: Autor).

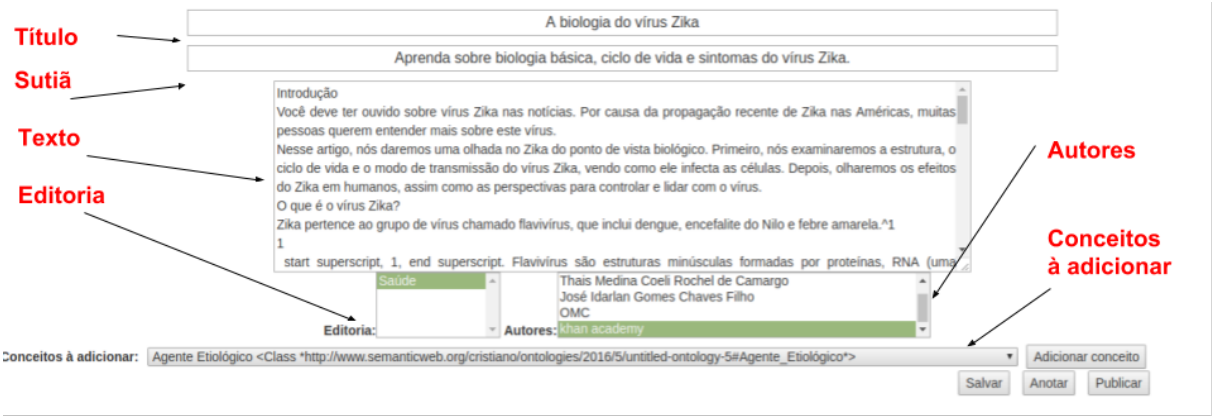

Figura 4.3: Tela de criação e edição de artigos.(Fonte: Autor).

4.2 Arquitetura da persistência de dados

É importante expor o modelo de banco de dados desse trabalho pois o mesmo contém todos os dados do sistema de gerenciamento de conteúdo e as triplas RDF das anotações semânticas em um mesmo banco relacional. Conjuntos de triplas RDF formam uma estrutura de grafo e portanto naturalmente deveriam ser armazenadas em bancos de dados baseados em grafo ou bancos não estruturados mas isso é uma opção como vemos nos exemplos presentes em [47]. As opções existentes seriam usar somente um banco baseado em grafo ou um banco não estruturado em que fosse possível representar todos os dados

Figura 4.4: Resultados da anotação de um artigo.(Fonte: Autor).

semânticos e não semânticos do protótipo, representar tudo em um banco relacional, ou construir uma solução híbrida. Dada a limitação de tempo, as ferramentas usadas e o foco do trabalho optou-se por construir um banco relacional para todos os dados persistidos e usados internamente no sistema, incluindo as anotações semânticas e essa decisão atende perfeitamente as necessidades. Entretanto para os dados oferecidos para sistemas semânticos exteriores foi possível disponibilizar os dados das anotações semânticas de cada artigo em um arquivo RDF com estrutura de grafo.

Pela Figura 4.5 pode-se observar que um artigo é constituído de título, texto, ”sutian”, editorias e autores. Cada artigo pode estar relacionado a várias editorias e a vários autores assim como uma dada editoria e um dado autor pode estar relacionado a vários artigos. Cada artigo possui zero ou mais artigos publicados relacionados. Os artigos publicados possuem o conteúdo do artigo que aparece na web(HTML), o arquivo RDF de anotações semânticas do mesmo e uma data de publicação. Cada artigo está relacionado a zero ou mais triplas que são construídas a partir da anotação semântica do texto do artigo. A tripla é representado por uma tabela no banco relacional e contém uma referência para seu artigo (sujeito) referência a um recurso (predicado) e um recurso (objeto). Cada recurso posssui um URI e um valor como `http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Agente_Etiológico` sendo o URI e Agente Etiológico o valor. Portanto o campo URI é a replicação de um identificador da Web Semântica no banco relacional interno do protótipo.

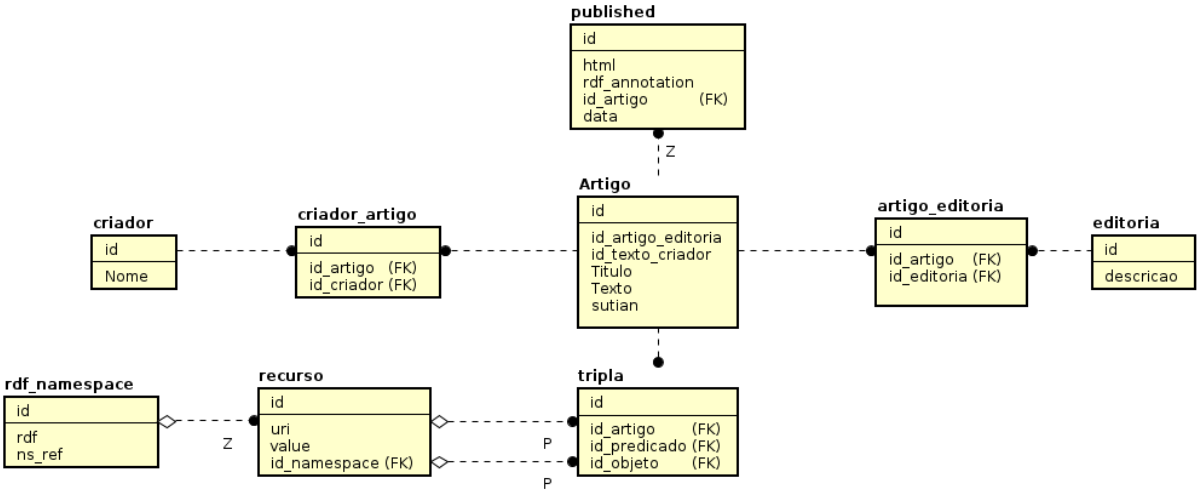

Figura 4.5: Modelo do banco de dados relacional.(Fonte: Autor).

Cada recurso pode estar relacionado ao seu namespace (contexto) que pode ser por exemplo uma ontologia da Web Semântica, ou seja, esse contexto pode também estar armazenado no banco local como é o caso da ontologia [7].

Esse modelo de banco foi inspirado nas soluções presentes em [47], que apresentam soluções para armazenar triplas RDF em bancos relacionais.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
   xmlns:ao="http://purl.org/ao/core/"
   xmlns:aof="http://purl.org/ao/foaf/"
   xmlns:ns1="http://cdn.rawgit.com/pav-ontology/pav/2.0/"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
>
  <rdf:Description rdf:nodeID="N976617b5b49e4bfcb0065a1274d8829d">
    <ns1:pav.owlcreatedOn rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2018-05-08T00:47:33.270623</ns1:pav.owlcreatedOn>
    <aof:annotatesDocument rdf:resource="www.article2.example">
    <ao:hasTopic rdf:resource="http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Microcefalia"/>
    <ns1:pav.owlcreatedB>Alexandre Vargas</ns1:pav.owlcreatedB>
    <rdf:type rdf:resource="http://purl.org/ao/core/Annotation"/>
  </rdf:Description>
  <rdf:Description rdf:nodeID="N37707bad9cc34d3e8e2347deb1a70911">
    <rdf:type rdf:resource="http://purl.org/ao/core/Annotation"/>
    <ao:hasTopic rdf:resource="http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Prurido"/>
    <ns1:pav.owlcreatedOn rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2018-05-08T00:47:33.271353</ns1:pav.owlcreatedOn>
    <ns1:pav.owlcreatedB>Alexandre Vargas</ns1:pav.owlcreatedB>
    <aof:annotatesDocument rdf:resource="www.article2.example"/>
  </rdf:Description>
```

Figura 4.6: Exemplo de arquivo com anotações semânticas de um artigo.(Fonte: Autor).

A Figura 4.6 traz um trecho do arquivo RDF que é armazenado na tabela *published* do banco relacional e é oferecido a sistemas semânticos externos. No início, dentro da tag rdf:RDF tem-se os *namespaces* ou contextos que são referenciados no arquivo, ou seja, que contém URIs que são referenciados no arquivo, que nessa caso são os propostos por [30] como, por exemplo, [32] e [28]. O restante do arquivo são tags rdf:Description cada uma relacionada a um conceito encontrado no texto do artigo, como por exemplo, Fase Viremia e Prurido. Desse modo, para cada uma desses conceitos anotados são registradas cinco

informações dentro da tag rdf:Description. O nodeID é um identificador único para esse item anotado dentro do documento. Na linha seguinte tem-se <aof:annotatesDocument rdf:resource=”www.article2example.com.br”/> onde aof:annotatesDocument significa que o predicado annotatesDocument do *namespace* aof está sendo referenciado e rdf:resource= "www.article2example.com.br"é o predicado,ou seja, temos que no nodeID "N976617b5b49 e4bfcb0065a1274d8829d"anota-se o artigo ”www.article2example.com.br”.

A linha <ns1:pav.owlcreatedOn rdf:datatype= `http://www.w3.org/2001/XMLSchema#date`>2018-05-08T00:47:33.270623
</ns1:pav.owlcreatedOn> também referência um *namespace*, um padrão de data e armazena um valor de data.

A linha <rdf:type rdf:resource=”`http://purl.org/ao/core/Annotation`”/> diz que o registro com determinado NodeID é do tipo referenciado em `http://purl.org/ao/core/Annotation`, ou seja, é uma anotação.

A linha <ns1:pav.owlcreatedB>Alexandre Vargas</ns1:pav.owlcreatedB> da mesma forma diz que o registro com o determinado NodeId foi criado por Alexandre Vargas, ou seja, ele foi quem anotou o termo ”Microcefalia” no artigo que está em `article2example.com.br`. A associação do artigo ao conceito ”Microcefalia” na Web Semântica é feita na linha <ao:hasTopic rdf:resource=”`semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Fase_Viremia`”,que diz exatamente que o artigo possui o conceito que está em:
`http://www.semanticweb.org/cristiano/ontologies/2016/5/untitled-ontology-5#Microcefalia`.

4.3 O algoritmo de anotação

A anotação semântica feita no sistema é uma anotação semi-automática baseada em ontologia, ou seja, é uma anotação feita por um ser humano com o apoio de um computador e uma ontologia. A Figura 4.7 apresenta um diagrama que mostra que o algoritmo de anotação recebe o texto e usando a ontologia produz uma lista de anotações sugeridas. Essa lista é filtrada e incrementada pelo autor do texto conforme sua percepção da semântica dos conceitos presentes no texto. Após a confirmação do autor as anotações são armazenadas no banco.

O primeiro passo do algoritmo é dividir o texto do artigo em uma lista de listas em que cada item da lista é um período do texto dividido em palavras na ordem em que as mesmas estão no período. Depois é construída uma lista com todos os conceitos a serem procurados no texto, nesse caso todos os presentes na ontologia do Zika [7]. Essa lista de conceitos é ordenada pelo número de palavras de cada conceito.

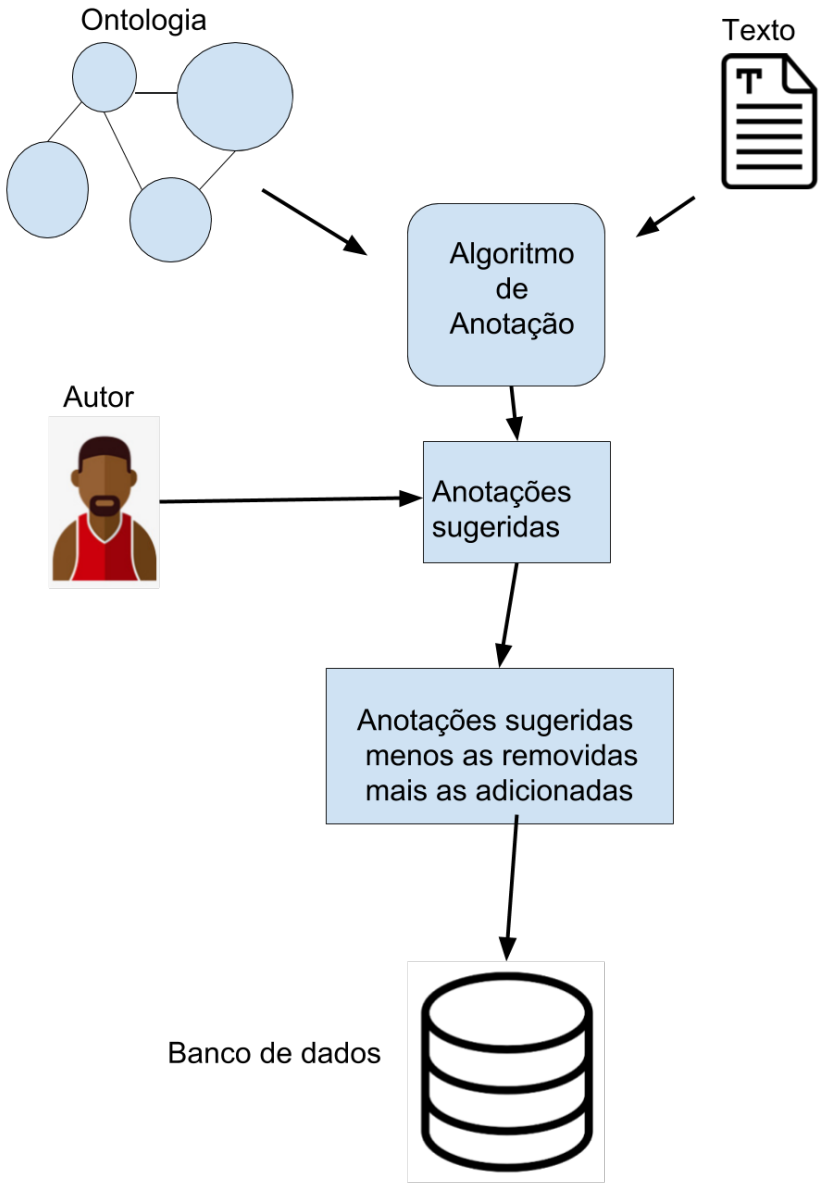

Figura 4.7: Processo de anotação semântica.(Fonte: Autor).

Para cada tamanho de conceito (número de palavras da representação textual do conceito) existente verifica-se se existem períodos com tamanho maior ou igual a aquele. Se não, remove-se todos os conceitos com esse determinado tamanho da lista. Se sim, procura-se esse determinado conceito em todos os períodos que possuem tamanho maior ou igual ao mesmo, caso seja encontrado em algum, o conceito é inserido em uma lista de conceitos para anotação e é removido da lista anterior, se não for encontrado nenhum o conceito também é removido da lista de conceitos que estão sendo procurados. O algoritmo segue até que a lista de conceitos esteja vazia.

Para cada conceito anotado também são anotados todos os nós pais na ontologia. Por

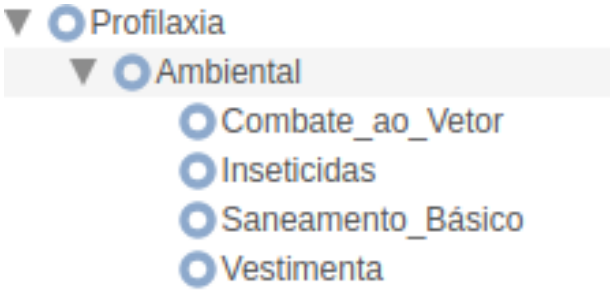

Figura 4.8: Exemplo de reificação em uma ontologia (Fonte: [7]).

exemplo, na Figura 4.8, Vestimenta, Saneamento Básico, Inseticidas ,Combate ao Vetor são instanciações de Profilaxia Ambiental e portanto um texto que fala de Vestimenta ou de Combate ao Vetor do Zika também fala de profilaxia ambiental do Zika, por exemplo.

O algoritmo, em pseudocódigo, é conforme o algoritmo 1.

Algorithm 1 Encontrar conceitos de uma lista em um texto.(Fonte: Autor)

```
procedure ENCONTRAR(lista lConceitos,texto)
    conceitos_encontrados ← lista_vazia()
    sentencas ← divide_em_sentenças(texto)
    conceitos_remanescentes ← lConceitos
    while nao_vazio(conceitos_remanescentes) do
        t ← tamanho(maior(i ∈ conceitos_remanescentes)
        m_conceitos ← lista(i ∈ lConceitos|tamanho(i) = t)
        conceitos_remanescentes ← lista(x|x ∉m_conceitos
        for sentence ∈ sentences do
            if tamanho(sentence) ≥ t) then
                for concept ∈ m_conceitos do
                    if concept ∈ sentence then
                        conceitos_encontrados.insere(conceito)
    retorna(conceitos_encontrados)
```

Para melhor compreensão traz-se em 4.3.1 um exemplo de execução do algoritmo a partir de um segmento de texto extraído de [48].

4.3.1 Texto de exemplo de entrada para o algoritmo 1

O principal meio de contágio do Zika vírus se dá através da picada do mosquito Aedes aegypti; o vírus Zika também é transmitido por relações sexuais, contato sanguíneo, leite materno e pelo líquido amniótico. A causa de microcefalia se dá quando a mãe está infectada e o vírus age perfurando a placenta chegando ao líquido amniótico infectando também o feto. Estudos apontam que o vírus destrói o tecido neuronal dos fetos. Nos casos de infecção nos primeiros 3 meses de gestação o feto tem mais chances de nascer com microcefalia.

O texto é então dividido nas quatro seguintes sentenças em forma de lista:

['O', 'principal', 'meio', 'de', 'contágio', 'do', 'Zika', 'vírus', 'se', 'dá', 'através', 'da', 'picada', 'do', 'mosquito', 'Aedes', 'aegypti', ';', 'o', 'vírus', 'Zika', 'também', 'é', 'transmitido', 'por', 'relações', 'sexuais', ',', 'contato', 'sanguíneo', ',', 'leite', 'materno', 'e', 'pelo', 'líquido', 'amniótico', '.']

['A', 'causa', 'de', 'microcefalia', 'se', 'dá', 'quando', 'a', 'mãe', 'está', 'infectada', 'e', 'o', 'vírus', 'age', 'perfurando', 'a', 'placenta', 'chegando', 'ao', 'líquido', 'amniótico', 'infectando', 'também', 'o', 'feto', '.']

['Estudos', 'apontam', 'que', 'o', 'vírus', 'destrói', 'o', 'tecido', 'neuronal', 'dos', 'fetos', '.']

['Nos', 'casos', 'de', 'infecção', 'nos', 'primeiros', '3', 'meses', 'de', 'gestação', 'o', 'feto', 'tem', 'mais', 'chances', 'de', 'nascer', 'com', 'microcefalia', '.']

A seguir temos uma lista com todos os conceitos que são reificações na ontologia como por exemplo o conceito *febre* que é uma reificação de sintoma. Todos os elementos dessa lista serão procurados nas sentenças conforme descrito no algoritmo.

[['Mielite'], ['Vestimenta'], ['Hidratação'], ['Relação', 'Sexual'], ['Repouso'], ['Notificação', 'Compulsória'], ['Saneamento', 'Básico'], ['Zika', 'Vírus'], ['Exantema'], ['Analgésico'], ['Zona', 'Endêmica'], ['Zika'], ['Prurido'], ['RT-PCR'], ['Anti-térmiico'], ['Conjuntivite'], ['Microcefalia'], ['Isolamento'], ['Febre'], ['Combate', 'ao', 'Vetor'], ['Dor', 'Articular'], ['Teste', 'de', 'Sangue'], ['Perguntas'], ['Usuário', 'de', 'Drogas'], ['Hemófilos'], ['Mosquito', 'Aedes', 'Egypt'], ['Alteração', 'Genética'], ['Picada', 'Mosquito'], ['Mal', 'Estar', 'Nevrálgico'], ['Zika', 'V'], ['Vertical'] ['Mal', 'Estar'], ['Vacina'], ['Mosquito'], ['Repelente'], ['Transfusão', 'de', 'Sangue'], ['Triagem'], ['Virus', 'Zika'], ['Evitar', 'Gravidez'], ['Dor', 'de', 'Cabeça'], ['Zika', 'Doença'], ['Habitante', 'de', 'Zona', 'Endêmica'], ['Gestante'], ['Síndrome', 'de', 'Guilan-Barré'], ['Rastreador'], ['População', 'Geral'], ['Inseticidas'], ['Líquido', 'Aminiótico']]

Desse modo, o primeiro conceito a ser procurado nas quatro sentenças é ['Habitante', 'de', 'Zona', 'Endêmica'] que não é encontrado e portanto descartado. Depois são procurados ['Combate', 'ao', 'Vetor'], ['Teste', 'de', 'Sangue'], ['Usuário', 'de', 'Drogas'], ['Mosquito', 'Aedes', 'Egypt'], ['Mal', 'Estar', 'Nevrálgico'], ['Transfusão', 'de', 'Sangue'], ['Dor', 'de', 'Cabeça'], ['Síndrome', 'de', 'Guilan-Barré']
que também não são encontrados e portanto descartados e parte-se para procurar:
['Relação', 'Sexual'], ['Notificação', 'Compulsória'], ['Saneamento', 'Básico'], ['Zika', 'Vírus'], ['Zona', 'Endêmica'], ['Dor', 'Articular'], ['Alteração', 'Genética'], ['Picada', 'Mosquito'], ['Zika', 'V'], ['Mal', 'Estar'], ['Virus', 'Zika'], ['Evitar', 'Gravidez'], ['Zika', 'Doença'], ['População', 'Geral'], ['Líquido', 'Amniótico'].
e então encontra-se ['Zika', 'Vírus'] na primeira sentença e passa-se a procurar por :
[['Mielite'], ['Vestimenta'], ['Hidratação'], ['Repouso'], ['Exantema'], ['Analgésico'], ['Zika'], ['Prurido'], ['RT-PCR'], ['Anti-térmiico'], ['Conjuntivite'], ['Microcefalia'], ['Isolamento'], ['Febre'], ['Perguntas'], ['Hemófilos'], ['Vertical'], ['Vacina'], ['Mosquito'], ['Repelente'], ['Triagem'], ['Gestante'], ['Rastreador'], ['Inseticidas']].
Desses encontra-se ['Zika'] e ['Mosquito'] na primeira sentença e ['Microcefalia'] na segunda finalizando-se assim o algoritmo.

Após isso são procurados todos os nós pais, na ontologia, das reificações encontradas para também serem anotados pois as reificações são instanciações de seus pais. A Microcefalia é um sintoma da fase viremia aguda da doença e portanto também podemos dizer que esse texto trata de um sintoma e da fase viremia do Zika pois a microcefalia é uma instanciação dos mesmos. Por fim a anotação é persistida no banco de dados.

4.4 A busca na base de dados relacional

O resultado da busca são artigos publicados no sistema. A mesma é feita a partir de seis campos, sendo dois com dados semânticos e quatro com dados comuns dos artigos que são texto, título, "sutian" (subtítulo), editorias e autores. Os dois campos semânticos são os campos com rótulo de *Conceitos* e *URIs* conforme a Figura 4.9. O campo *Conceitos* diz respeito ao campo que armazena uma representação textual de um conceito da ontologia no banco relacional e é mais especificamente o campo *value* da tabela *recurso* mostrada na Figura 4.5, ou seja, procura os recursos da Web Semântica que estão relacionados a aquele artigo por meio de uma representação textual simples. O campo *URIs* realiza também busca na tabela *recurso* mas no campo *uri*, ou seja, realiza busca na representação real daquele recurso ou conceito na Web semântica. Os campos *Editorias* e *Autores* permitem filtrar os artigos por seus autores e editorias. Os demais campos permitem filtrar os artigos pelos próprios campos da tabela *Artigo*.

Figura 4.9: Campos da tela de busca de artigos publicados.(Fonte: Autor).

Para cada campo é executada uma consulta diferente no banco e depois é gerada a união dessas consultas como resultado. Cada campo aceita uma expressão de busca que pode conter parênteses, aspas, operador OR (|) ou operador AND (&) em notação infixa. Alguns exemplos de busca são os apresentados na Figura 4.10 e na Figura 4.11.

Na Figura 4.10 apresenta-se um exemplo de consulta no qual busca-se os artigos que falam sobre febre ou tratamento ou que contenham a palavra dengue no título.

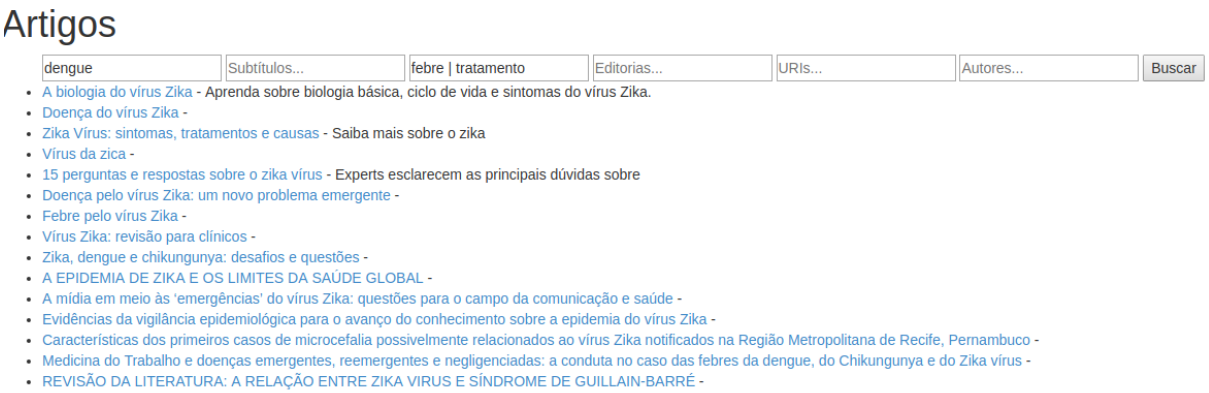

Figura 4.10: Exemplo de busca.(Fonte: Autor).

Na Figura 4.11 apresenta-se um exemplo de consulta no qual busca-se os artigos que tenham as palavras Zika e doença ou a palavra febre no título.

Figura 4.11: Exemplo de busca.(Fonte: Autor).

Em suma, a camada que realiza cada consulta a partir de uma expressão de busca realiza um pré-processamento da expressão montando uma lista de operadores, parênteses e operandos em que cada posição da lista é um operador ou parênteses ou palavra ou palavras quando entre aspas, todos na mesma posição infixa na qual foram informados no campo de busca. Após isso a expressão é passada para a notação posfixa e cada operando

é substituído por um conjunto de artigos aos quais se remetem aquela palavra ou palavras entre aspas. Por exemplo, ao pesquisar a palavra "Vitor" no campo autores tem-se todos os artigos cuja palavra 'Vitor' está contida no nome de seus autores. "Vitor Silva" e "Vitor Laerte", por exemplo. Depois a expressão em notação posfixa na qual os operandos são conjuntos de artigos é processada realizando intersecção quando o operando for "&" e união quando for "|". Outro exemplo seria pesquisar por "Vitor Silva"no título e o retorno seria todos os artigos que contenham a palavra "Vitor" ou a palavra "Silva" no título. Ou mesmo pesquisar por "Vitor Silva" e ter como retorno os que tenham "Vitor Silva" no título.

4.5 Abordagens para inferência de relacionamento semântico entre os textos

Duas possíveis abordagens para o relacionamento entre textos foram propostas aqui. A primeira, em suma , faz com que um texto B de um conjunto C de textos seja o mais relacionado a um outro texto A se B é o texto cuja intersecção entre os conceitos anotados de B e e A seja a maior possível dentre todos os outros textos de C. Portanto se deseja-se saber, a partir dessa abordagem, quais os cinco textos de um conjunto C mais relacionados a um texto A, obtém-se os cinco textos com o maior tamanho de intersecção de seus conceitos anotados com os conceitos anotados do artigo A ordenados pelo tamanho dessa intersecção.

A segunda, um pouco mais complexa, usa a estrutura das ontologias que contém os conceitos relacionados aos textos para extrair uma métrica para o relacionamento entre dois textos. Em suma, essa métrica é a quantidade de conceitos irmãos que existem entre os conceitos anotados de dois artigos A e B incluindo o próprio conceito, ou seja, a quantidade de conceitos do artigo A que possuem o mesmo pai em alguma ontologia que um conceito de B. Por exemplo, observando a Figura 4.12 percebe-se que a intersecção entre os conceitos de dois artigos tais que um fala do Vírus Zika e do Zika Doença é vazia. Entretanto os dois artigos estão falando de Agente Etiológico, que é o pai dos dois conceitos anteriores na ontologia e portanto os dois artigos estão falando de assuntos relacionados em um certo nível de abstração e essa segunda abordagem leva isso em consideração.

Os resultados obtidos nos dois casos dependem bastante da massa de dados do sistema. Não é trivial extrair uma métrica bem embasada sobre eficiência e característica das duas abordagens sem um estudo detalhado com uma grande massa de dados, e isso não foi o foco deste trabalho. Entretanto para fins de exemplificar o funcionamento do sistema apresenta-se aqui alguns resultados qualitativos obtidos dos dois algoritmos a partir de

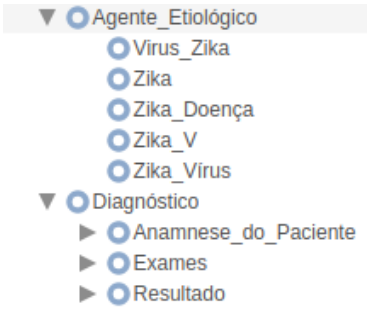

Figura 4.12: Exemplo de conceitos irmãos na ontologia (Fonte: [7]).

uma massa pequena e restrita ao domínio do Zika Vírus apresentado em [7]. Foram inseridos no banco de dados artigos jornalísticos e científicos que tratam do Zika conforme a Tabela 4.1

Tabela 4.1: Tipos de artigos inseridos no sistema

Tipo de artigo	Quantidade
Jornalístico	14
Científico	16
Total	30

Para a primeira abordagem, com textos restritos a um mesmo domínio, que é o caso da massa de teste ao qual estamos submetendo os dois algoritmos, uma característica fica clara a partir de simples observação: Textos que referenciam muitos conceitos diferentes do domínio acabam sendo relacionados a muitos outros textos pois as chances de que a intersecção de seus conceitos com os de outros ser grande são maiores. Um exemplo é o artigo *Medicina do Trabalho e doenças emergentes, reemergentes e negligenciadas: a conduta no caso das febres da dengue, do Chikungunya e do Zika* [2] está entre os cinco mais relacionados a dezoito de outros vinte artigos já cadastrados no momento dessa avaliação e contém trinta e nove dos quarenta e cinco conceitos presentes na ontologia.

Na segunda abordagem, o artigo [2] continuou aparecendo em dezessete de outros vinte artigos, o que é lógico pois um artigo que possui 39 dos 45 conceitos mapeados do domínio também terá mais conceitos irmãos ou iguais aos outros artigos do domínio. No geral os cinco primeiros textos relacionados não se alteraram muito. Por regra, mudaram de ordem e se modificaram em dois ou três como foi o caso de [2](Figura 4.13 e Figura 4.14) e [3] (Figura 4.15 e Figura 4.16).

Uma forma de gerar resultados mais precisos sobre as duas abordagens seria, com o auxílio de um especialista do domínio, popular o sistema com uma massa maior de textos do domínio de forma que grupos de textos pertencessem a subdomínios em comum dentro

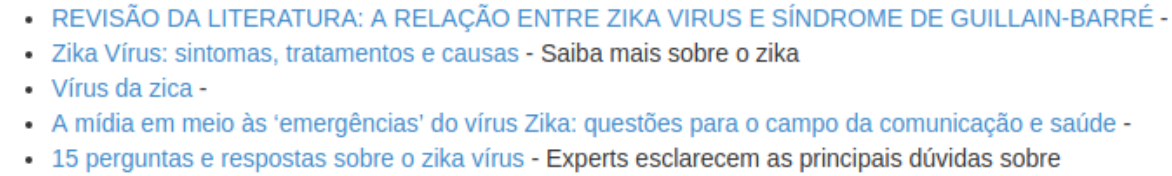

Figura 4.13: Textos relacionados ao artigo [2] na primeira abordagem.(Fonte: Autor).

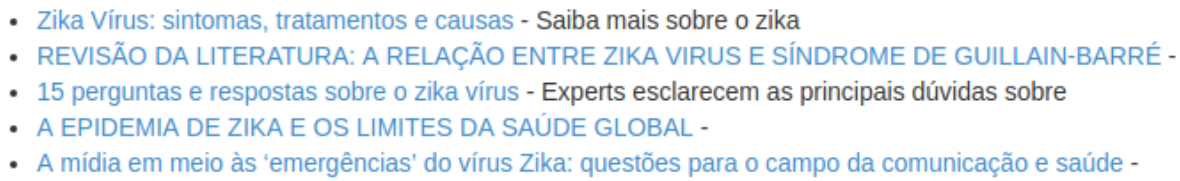

Figura 4.14: Textos relacionados ao artigo [2] na segunda abordagem.(Fonte: Autor).

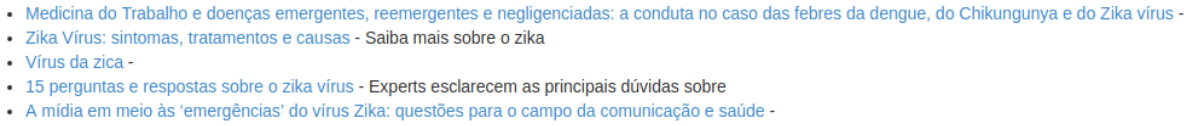

Figura 4.15: Textos relacionados ao artigo [3] na primeira abordagem.(Fonte: Autor).

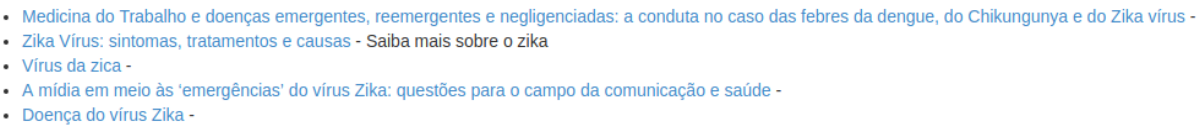

Figura 4.16: Textos relacionados ao artigo [3] na segunda abordagem.(Fonte: Autor).

do domínio maior e então extrair métricas do quão bem os algoritmos relacionam textos dentro dos subdomínios e do domínio que os contém.

Outra observação importante é a de que os artigos científicos, em média, possuem mais triplas geradas pela anotação automática, ou seja, sugestões de anotações feitas pelo sistema, conforme mostra a Tabela 4.2. Isso pode ser interpretado como uma possível superficialidade dos artigos jornalísticos, ou seja, os artigos científicos em média citam mais termos do domínio da ontologia do zika do que os artigos jornalísticos. O desvio padrão das médias de ambos os tipos de artigos mostra que as amostras são bastante variadas com respeito ao número de triplas associadas a cada artigo.

Tabela 4.2: Informações sobre a quantidade média de triplas RDF associadas a cada tipo de artigo

Tipo de artigo	Quantidade média de triplas	Desvio padrão da média de triplas
Científico	16,93	9,3
Jornalístico	11,07	8,06

4.5.1 Conclusão

O foco foi alcançar os objetivos propostos para o artefato sem explorar os aspectos de aparência das telas mas construindo uma interface que atendesse minimamente às necessidades de forma que o artefato ainda não é acabado mas é completamente funcional e testável. Nesse capítulo foram expostos alguns dos passos necessários para o alcance dos objetivos assim como demonstrações de uso do artefato,ou seja, a arquitetura da persistência de dados, o algoritmo de anotação semântica, o algoritmo e o protótipo da tela de busca na base de dados relacional, as abordagens para inferência de relacionamento semântico entre os textos, o protótipo da tela de criação,edição e anotação dos artigos. Além disso foi exposta a constatação com respeito a diferença na quantidade média de triplas existentes em artigos jornalísticos e científicos, conforme a tabela 4.2. No próximo Capítulo os objetivos alcançados são apresentados de forma geral e relacionados com a implementação. Também são apresentados possíveis trabalhos futuros.

Capítulo 5

Conclusões

Alguns gargalos presentes em um CMS não semântico são busca, relacionamento automático entre os textos e comunicação com sistemas semânticos externos como, por exemplo, um sistema que faça busca semântica nos textos do sistema aqui apresentado a partir da interface RDF oferecida. Um exemplo desses gargalos seria que quando se busca por "Barcelona" em uma base de dados pode-se ter como retorno textos sobre o time de futebol Barcelona e sobre a cidade Barcelona. A busca proposta neste trabalho resolve esse problema permitindo buscas com URIs assim como os algoritmos de relacionamento dos textos a partir de seus URIs completando então o objetivo 3 da seção 1.1.2. A interface semântica oferecida para sistemas externos contendo o HTML e um RDF com as anotações semânticas de cada artigo oferece uma estrutura bem definida que reusa recursos da Web semântica e que pode ser perfeitamente consumida por qualquer agente exterior cumprindo então o objetivo 4 da seção 1.1.2. A confiabilidade das anotações semânticas produzidas semi-automaticamente é garantida pelo fato de que além da geração automática das mesmas, o humano que está anotando pode retirar ou adicionar anotações conforme sua compreensão do texto. A anotação de conceitos que não foram diretamente citados no texto mas que estão presentes devido a sua relação com algum conceito anota também é feita. Por exemplo, se o texto fala de febre do Zika então ele também fala de um sintoma do Zika e essa informação também é anotada, ou seja, o suporte automático feitos as anotações semânticas é providencial e completa com qualidade o requisito de possibilitar anotações semânticas semi-automáticas presentes no objetivo 1 da seção 1.1.2. Para o relacionamento entre textos foram aqui apresentadas duas opções viáveis. Uma das opções leva apenas em consideração os conceitos anotados de um artigo com relação aos anotados de outro e a outra opção também leva em consideração a generalização dos conceitos, ou seja, se conceitos diferentes são instanciações de um mesmo conceito então eles são relacionados. Essas duas abordagens implementadas atendem completamente ao segundo objetivo específico. Tendo isso em vista pode-se afirmar que por fim tem-se um

protótipo de CMS semântico funcional, porque é testável e implantável em um ambiente real de uso e evolutivo porque sua mesma estrutura pode evoluir para um CMS real completando assim o último objetivo específico proposto.

5.1 Trabalhos Futuros

5.1.1 Ontologia de domínio com reuso

Ontologias podem se relacionar com outras ontologias. Isso ainda não é o caso da ontologia apresentada em [7]. O conceito dor de cabeça, por exemplo poderia apenas ser uma referência para uma outra ontologia que fala sobre dores ou dor de cabeça em específico. Além disso outras ontologias de domínio podem ser inseridas no sistema para embasar a criação das anotações semânticas.

5.1.2 Anotação semântica

A parte automática da construção da anotação semântica nesse trabalho é feita simplesmente por comparar a representação textual de um conceito com segmentos dos textos. É possível incrementar esse processo adicionando técnicas de processamento de linguagem natural, aprendizagem de máquina e outros como é feito em [41], por exemplo.

5.1.3 Persistência dos dados e busca semântica

Aumentar o uso de bancos de dados não estruturados no sistema pode abrir novos horizontes, afinal a Web Semântica é uma estrutura de grafo com dados listados de forma não estruturada e não uma estrutura relacional com dados tipados. Essa evolução possibilitaria consultas SPARQL, ou seja, construir consultas que tenham como entrada dados internos e externos ao sistema, pois todos seriam dados disponíveis na Web Semântica.

5.1.4 Avaliação

Implantar e avaliar os ganhos da utilização do artefato em um sistema de informação real. Avaliar os algoritmos de relacionamento de texto formalmente com especialistas no domínio em que as anotações estão sendo criadas e uma base de dados maior e construída a partir de um embasamento claro para avaliação como , por exemplo, o apresentado em [35] .

Referências

[1] Aline Dresch, José Antonio Valle Antunes Júnior, Daniel Pacheco Lacerda: *Design science research: método de pesquisa para avanço da ciência e tecnologia.* bookman, Porto Alegre, BR, 2015. ix, 3, 22, 26

[2] Pustiglione, Marcelo: *Medicina do trabalho e doenças emergentes, reemergentes e negligenciadas: a conduta no caso das febres da dengue, do chikungunya e do zika vírus / biblioteca virtual em saúde.* `http://pesquisa.bvsalud.org/cvsp/resource/pt/lil-779356?lang=pt`. Acessado em 07/06/2018. ix, 39, 40

[3] Jose Idarlan Gomes Chaves Filho, Isabella de Lara Aires Reis, Adrieli dos Santos França Denise da Costa Boamorte Cortela: *Revisão da literatura: a relação entre zika vírus e síndrome de guillain-barré.* `https://periodicos.unemat.br/index.php/revistamedicina/article/view/1365`. Acessado em 07/06/2018. x, 39, 40

[4] Tim Berners-Lee, James Hendler, Ora Lassila: *The semantic web a new form of web content that is meaningful to computers will unleash a revolution of new possibilities.* Scientific American, 2001. 1, 7, 8

[5] Edison Ishikawa, Benedito Medeiros Neto, George Ghinea: *Newsroom 3.0: Managing technological and media convergence in contemporary newsrooms.* Relatório Técnico, 2018. 2, 3, 14, 15

[6] Journalists. ICFJ, International Center for: *A study of technology in newsroom.* 2, 3

[7] Edgard Costa Oliveira, Edison Ishikawa, George Ghinea Thabata Hellen Granja Marcos Nunes Lucas Hiroshi Hironouchi Rafael Batista Menegassi Luciano Gois Daniel Rodriguez: *Designing an ontology-based zika virus news authoring environment for the semantic web.* roceedings of the 8th International Conference on Management of Digital EcoSystems, novembro 2016. 2, 25, 27, 31, 32, 34, 39, 43

[8] Raul Sidnei Wazlawick: *Metodologia de Pesquisa para Ciência da Computação.* Elsevier, 2009. 4

[9] Seiji Isotani, Ig Ibert Bittencourt: *Dados Abertos Conectados.* Novatec, 2015. 5, 6, 9, 12

[10] Thomas Gruber: *A translation approach to portable ontology specifications.* Knowledge Acquisition June 1993, Pages 199-220, 5. 5

[11] Rudi Studera , Richard Benjamin, Dieter Fensela: *Knowledge Engineering, Principles and Methods*, volume 25. 1998. 5

[12] Peng Wang, Bao-wen Xu, Jian-jiang Lu Da-zhou Kang Yan-hui Li: *A novel approach to semantic annotation based on multi-ontologies*. Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference, 2004. 5

[13] Andreza Regina Lopes da Silva, Michele Andréia Borges, Maria Cristina Pfeiffer Fernandes Viviane Sartori Fernando José Spanhol: *Ontologia como representação do conhecimento: aplicação no curso de formação continuada em tecnologias educacionais na web*. 2014. 6

[14] Nicola Guarino: *Formal ontology and information systems*. Proceedings of FOIS'98, 1998. 6

[15] Stanford University: *https://protege.stanford.edu/ontologies/pizza/pizza.owl*. `https://protege.stanford.edu/ontologies/pizza/pizza.owl`. Acessado em 07/06/2018. 7

[16] Mijung Kim, Jake Cobb, Mary Jean Harrold Tahsin Kurc Alessandro Orso Joel Saltz Andrew Post Kunal Malhotra Shamkant Navathe: *Efficient regression testing of ontology-driven systems*. Proceedings of the 2012 International Symposium on Software Testing and Analysis, 2014. 7

[17] James Hendler: *Is there an intelligent agent in your future?* `www.nature.com/nature/webmatters/agents/agents.html`, acesso em 11 Mar. 1999, Acessado em 07/06/2018. 8

[18] *W3c semantic web frequently asked questions*. `https://www.w3.org/2001/sw/SW-FAQ#swgoals`, Acessado em 07/06/2018. 9

[19] Karin Koogan Breitman: *Web Semântica a Internet do futuro*. LTC, 2005. 9

[20] W3c: *Rdf primer*. `https://www.w3.org/TR/rdf-primer/`. Acessado em 07/06/2018. 9

[21] Stanford University: *Web protegé*. `https://protege.stanford.edu/`. Acessado em 07/06/2018. 9

[22] W3C: *Web ontology language*. `https://www.w3.org/OWL/`, Acessado em 07/06/2018. 10

[23] W3C: *Owl web ontology language overview*. `https://www.w3.org/TR/owl-features/`, Acessado em 07/06/2018. 10

[24] W3C: *Sparql query language for rdf*. `https://www.w3.org/TR/rdf-sparql-query/`, Acessado em 07/06/2018. 10, 11

[25] Cambridge University: *Owl web ontology language overview*. `https://www.cambridgesemantics.com/blog/semantic-university/learn-sparql/sparql-vs-sql/`, Acessado em 07/06/2018. 11

[26] Vipin Kumar, Archana Kumar, Kumar Abhishek: *A comprehensive comparative study of sparql and sql*. International Journal of Computer Science and Information Technologies, 2, 2011. 11

[27] Eyal Oren, Knud Hinnerk Moller, Simon Scerri,Siegfried Handschuh Michael Sintek: *What are semantic annotations?* janeiro 2006. 11

[28] Dan Brickley, Libby Miller: *Foaf vocabulary specification.* `http://xmlns.com/foaf/spec/`. Acessado em 07/06/2018. 12, 15, 21, 31

[29] *Google acadêmico.* `https://scholar.google.com.br/`, Acessado em 07/06/2018. 13, 23

[30] Paolo Ciccarese, Marco Ocana, Leyla Jael Garcia Castro Sudeshna Das Tim ClarkE-mail: *An open annotation ontology for science on web 3.0.* Journal of Biomedical Semantics, 2(2), 2011. 15, 16, 21, 31

[31] W3C: *Xml, xlink and xpointer.* `https://www.w3schools.com/xml/xml_xlink.asp`. Acessado em 07/06/2018. 15

[32] Paolo Ciccarese, Marco Ocana, Leyla Jael Garcia Castro Sudeshna Das Tim ClarkE-mail: *Ao.* `http://annotation-ontology.googlecode.com/svn/trunk/`, Acessado em 07/06/2018. 15, 21, 31

[33] Paolo Ciccarese, Stian Soiland-Reyes, Khalid Belhajjame Alasdair JG Gray Carole Goble and Tim Clark: *Pav ontology: provenance, authoring and versioning.* Acessado em 07/06/2018. 15

[34] Celso Araujo Fontes, Maria Cláudia Cavalcanti, Ana Maria Moura: *An ontology-based reasoning approach for document annotation.* Semantic Computing (ICSC),IEEE Seventh International Conference, 2013. 16, 18

[35] Maryam Hazman, Samhaa El-Beltagy, Ahmed Rafea: *An ontology based approach for automatically annotating document segments.* International Journal of Computer Science Issues,, 9(2), 2012. 17, 18, 21, 43

[36] Kavita Ganesan: *A brief note on stop words for text mining and retrieval.* outubro 2014. 17, 19

[37] Kiril Simov, Petya Osenova, Alexander Simov Anelia Tincheva Borislav Kirilov: *A system for a semi-automatic ontology annotation.* dezembro 2016. 17

[38] Princeton University: *Wordnet | a lexical database for english.* `https://wordnet.princeton.edu/`. Acessado em 07/06/2018. 19

[39] Che-Yu Yang, Hua-Yi Lin: *Semantic annotation for the web of data: An ontology and rdf based automated approach.* Journal of Convergence Information Technology, 6(4), 2011. 19, 21

[40] Ming Che Lee, Hui Hui Chen, Yu Sheng Li: *Fca based concept constructing and similarity measurement algorithms.* IJACT, 3(1):97–105, 2011. 19, 20

[41] Quratulain Rajput, Sajjad Haider: *Bnosa: A bayesian network and ontology based semantic annotation framework.* Web Semantics: Science, Services and Agents on the World Wide Web, 9(2), 2011. 20, 21, 43

[42] Judea Pearl: *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988. 20

[43] Django Project: *Django*. `https://www.djangoproject.com/`, Acessado em 07/06/2018. 24

[44] RDFLib Project: *Rdflib is a python library for working with rdf, a simple yet powerful language for representing information.* `https://github.com/RDFLib/rdflib`, Acessado em 07/06/2018. 24

[45] RDFLib Project: *Ontospy*. `https://github.com/lambdamusic/OntoSpy/wiki`, Acessado em 07/06/2018. 24

[46] Edgard Costa Oliveira. Tese de Doutoramento. 27

[47] Stanford University: *Storing rdf in a relational database*. `http://infolab.stanford.edu/~melnik/rdf/db.html`. Acessado em 07/06/2018. 29, 31

[48] Luiz de Oliveira Alves: *Zika vírus - infoescola*. `https://www.infoescola.com/doencas/zika-virus/`. Acessado em 07/06/2018. 34

Apêndice A

Utilização do sistema

A.1 Configuração do servidor de aplicação para sistemas linux derivados do Ubuntu

Após obter permissão de acesso ao repositório do trabalho,abra o terminal e execute o comando:

- git clone https://github.com/edisonik/newsroomFramework.git
- cd newsroomFramework
- sudo apt-get update
- xargs -a dependencies.txt sudo apt-get install
- virtualenv –p /usr/bin/python3 virtualenv
- source virtualenv/bin/activate
- pip install -r python_requirements.txt
- mysql -u root -p
- CREATE USER 'seu_usuario'@'localhost' IDENTIFIED BY 'sua_senha';
- GRANT ALL PRIVILEGES ON *.* TO 'seu_usuario'@'localhost' WITH GRANT OPTION;
- CREATE DATABASE cms CHARSET utf8;
- Em settings.py modifique as entradas USERNAME e PASSWORD do dicionário DATABASES para o usuário e senha escolhidos ,respectivamente.
- python manage.py migrate

É importante salientar que distribuições de sistema operacional podem não se encaixar perfeitamente neste passo à passo. Fica ao leitor a tarefa de adequação a depender de cada caso.

A.2 Testes e acesso à aplicação

- source virtualenv/bin/activate
- python manage.py runserver
- Acesse http://127.0.0.1:8000/menu/ para navegar na aplicação

A.3 Sistema administrador do Django

É um sistema de gerenciamento de conteúdo oferecido pelo django que fornece interfaces para vizualização, inserção e atualização dos dados de uma base de dados, nesse caso aos dados de Autores, Editorias e Artigos da base de dados definida na Figura 4.5.

- source virtualenv/bin/activate
- Crie um usuário com o comando: python manage.py createsuperuser
- python manage.py runserver
- Acesse http://127.0.0.1:8000/admin/